Reminder

- HW 1: Today, must be submitted by a hard copy to TA's office by 6 pm
- HW 2: Due April 22 (Tue)
- Midterm Review (April 21)
- Midterm (April 24)

Agenda

- Association statistic
- Statistical Power
- Association Power
- Relative risk example
- Indirect Association
- Multiple hypothesis testing
- HW1
- HW2

First, Notation (Very important!)

$\mathcal{N}$ = the number of total individuals

$\frac{\mathcal{N}}{2}$ = the number of case individuals

$\frac{\mathcal{N}}{2}$ = the number of control individuals

$\hat{p}_A^+$ = Observed case frequency (frequency from data)

$\hat{p}_A^-$ = Observed control frequency (frequency from data)

p_A^+ = True case frequency (never known)

p_A^- = True control frequency (never known)

Allele frequency and its distribution

- We have N/2 cases and N/2 controls
- Each individual has **2** chromosomes
- So, we have **N** case chromosomes and **N** control chromosomes
- We know the following

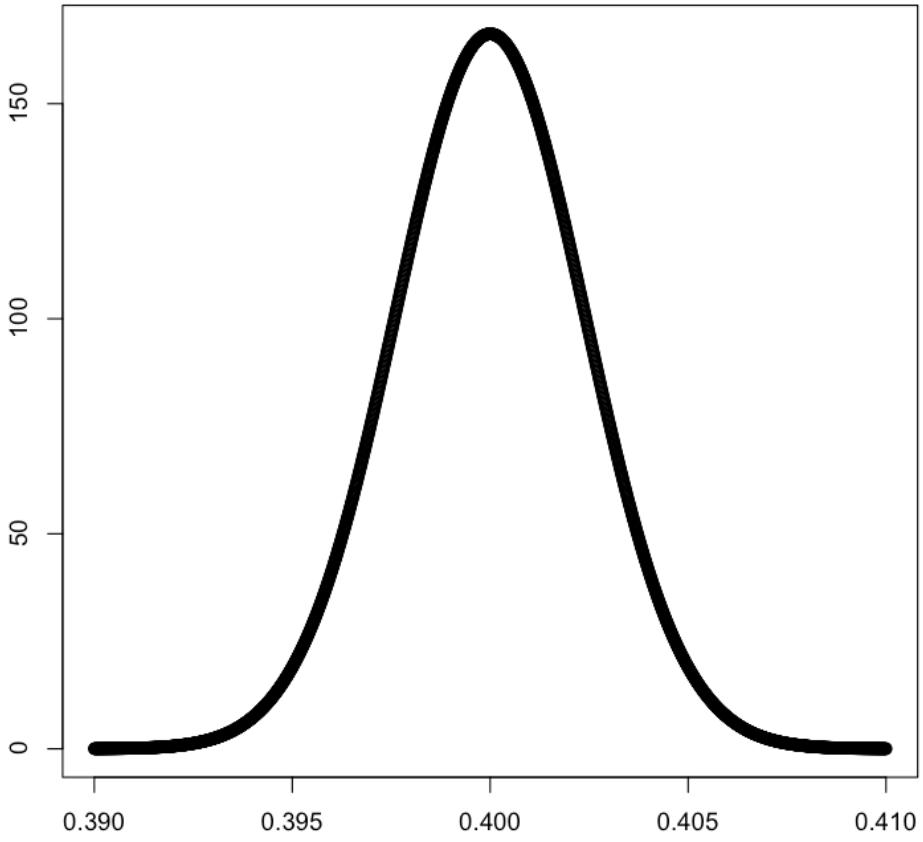
$$\hat{p}_A^+ \sim N(p_A^+, p_A^+(1-p_A^+)/N)$$

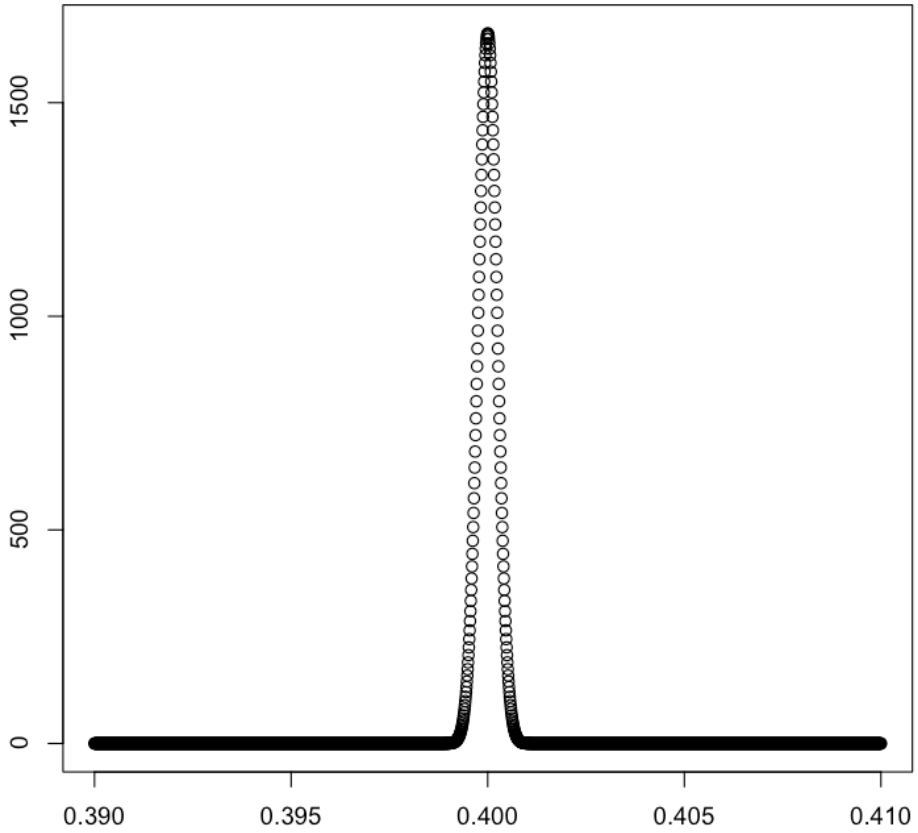
$$\hat{p}_A^- \sim N(p_A^-, p_A^-(1-p_A^-)/N)$$

Mean Variance

- What this says is that the frequency we observe from the data approaches to the true frequency when N is large (because variance is small)

Allele frequency and its distribution





$p_A^+ = 0.4$ and $\mathcal{N} = 100$

$p_A^+ = 0.4$ and $\mathcal{N} = 1000$

A difference in allele frequency

- We have the following rule for the normal distribution

$$X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2), \text{ then}$$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Not subtract! Add! (Many mistakes in midterm)

- Then, let's take a difference in observed frequency between cases and controls

$$\hat{p}_A^+ \sim N(p_A^+, p_A^+(1-p_A^+)/N) \quad \hat{p}_A^- \sim N(p_A^-, p_A^-(1-p_A^-)/N)$$

$$\hat{p}_A^+ - \hat{p}_A^- \sim N(p_A^+ - p_A^-, (p_A^+(1-p_A^+) + p_A^-(1-p_A^-))/N)$$

One approximation

- We make the following approximation to simplify our variance term

$$p_A^+(1-p_A^+)+p_A^-(1-p_A^-)\approx 2p_A(1-p_A)$$

$$\text{where } p_A=\frac{p_A^+ + p_A^-}{2}$$

- Then, our variance becomes

$$\hat{p}_A^+ - \hat{p}_A^- \sim N(p_A^+ - p_A^-, 2p_A(1-p_A)/N)$$

Normalization (divide by standard deviation)

- We have the following (another) rule for the normal distribution

$$X \sim \mathcal{N}(\mu_X, \sigma_x^2),\ aX \sim \mathcal{N}(a\mu_X, a^2\sigma_x^2),\ \frac{X}{a} \sim \mathcal{N}\left(\frac{\mu_X}{a}, \frac{\sigma_x^2}{a^2}\right)$$

- We want our variance to be 1. How?
  - Divide whole thing by the standard deviation (square root of variance)

$$\hat{p}_A^+ - \hat{p}_A^- \sim \mathcal{N}(p_A^+ - p_A^-, 2p_A(1-p_A)/N)$$

$$\text{Standard deviation} = \sqrt{(2p_A(1-p_A))/N}$$

$$\frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{(2\hat{p}_A(1-\hat{p}_A))/N}} \sim \mathcal{N}\left(\frac{p_A^+ - p_A^-}{\sqrt{(2p_A(1-p_A))/N}}, \frac{2p_A(1-p_A)/N}{2p_A(1-p_A)/N}\right)$$

Normalization (divide by standard deviation)

$$\frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{(2\hat{p}_A(1-\hat{p}_A))/N}} \sim \mathcal{N}\left(\frac{p_A^+ - p_A^-}{\sqrt{(2p_A(1-p_A))/N}}, \frac{2p_A(1-p_A)/N}{2p_A(1-p_A)/N}\right)$$

$$= \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{(2\hat{p}_A(1-\hat{p}_A))/N}} \sim \mathcal{N}\left(\frac{p_A^+ - p_A^-}{\sqrt{(2p_A(1-p_A))/N}}, 1\right)$$

Hat!

No Hat!

$$\hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} \qquad p_A = \frac{p_A^+ + p_A^-}{2}$$

Association Statistic

$$S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{(2\hat{p}_A(1-\hat{p}_A))/N}} \sim \mathcal{N}\left(\frac{p_A^+ - p_A^-}{\sqrt{(2p_A(1-p_A))/N}}, 1\right)$$

If $p_A^+ - p_A^- = 0$ (Null Hypothesis), $S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim \mathcal{N}(0,1)$

When computing p-value of association statistic

If $p_A^+ - p_A^- \neq 0$ (Alt Hypothesis), $S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim \mathcal{N}\left(\lambda_A\sqrt{N}, 1\right)$

where $\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}$, noncentrality-parameter is $\lambda_A\sqrt{N}$

When computing association power

Association Statistic – Computing p-value

- 100 cases, 100 controls, significance threshold $\alpha = 0.05$
- Observe 130 A's in cases and 110 A's in controls

$$\hat{p}_A^+ = \frac{130}{200} = .65 \quad \hat{p}_A^- = \frac{110}{200} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .6$$

$$S_A = \frac{\hat{p}^+{}_A - \hat{p}^-{}_A}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} = \frac{.65 - .55}{\sqrt{2/200}\sqrt{.6(1-.6)}} = 2.04$$

- Is this statistic (S_A) significant given the significance threshold?

Association Statistic – Computing p-value

- One way is to find whether S_A is in the tail of the normal distribution
- First, we find where the significance threshold (α=0.05) is on the standard normal distribution

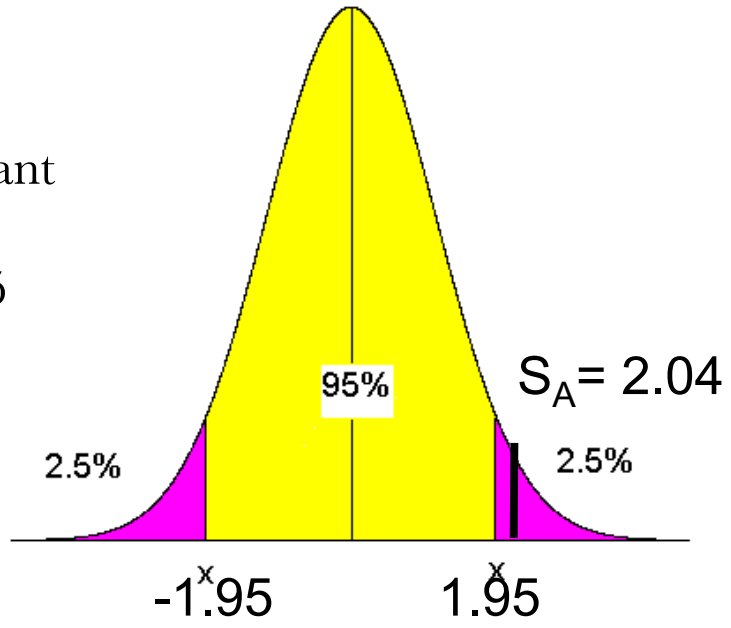
qnorm computes the value of x when we know $\Pr(X \le x)$, inverse of CDF

$$\Phi^{-1}(\alpha/2) = \Phi^{-1}(0.025) = \text{qnorm}(0.025) = -1.95$$

$$\Phi^{-1}(1-0.025) = \Phi^{-1}(0.975) = \text{qnorm}(0.975) = 1.95$$

If $S_A < \Phi^{-1}\left(\frac{\alpha}{2}\right)$ or $S_A > -\Phi^{-1}\left(\frac{\alpha}{2}\right)$, then significant

In this case, check whether $S_A < -1.95$ or $S_A > 1.95$



Association Statistic – Computing p-value

- Another way is to use pnorm in R
- pnorm computes Pr ($X \leq S_A$) or Pr($X > S_A$)
- If S_A is positive, p-value = 2*pnorm(2.04, lower.tail=F) = 2*0.021 = 0.042
- If S_A is negative, p-value = 2*pnorm(-2.04) = 2*0.021 = 0.042
- If p-value is less than the significance threshold (α=0.05), it is significant

Association Statistic – Another example

- 1000 cases, 1000 controls, significance threshold $\alpha = 0.05$
- Observe 1200 A's in cases and 1100 A's in controls

$$\hat{p}_A^+ = \frac{1200}{2000} = .6 \quad \hat{p}_A^- = \frac{110}{2000} = .55 \quad \hat{p}_A = \frac{\hat{p}_A^+ + \hat{p}_A^-}{2} = .575$$

$$S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} = \frac{.6 - .55}{\sqrt{2/2000}\sqrt{.575(1-.575)}} = 3.19$$

- Is this statistic (S_A) significant given the significance threshold?

$$\text{Check if } S_A < \Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ or } S_A > -\Phi^{-1}\left(\frac{\alpha}{2}\right)$$

$$\text{In this case, check whether } S_A < -1.95 \text{ or } S_A > 1.95$$

- p-value = 2*pnorm(3.19, lower.tail=F) = 2*0.00071 = 0.00142
- This p-value (0.00142) is less than significance threshold (0.05), so significant

Statistical Power

- So far, we assumed there is no effect (a fair coin, SNP is not associated) and computed a p-value
- Now, let's assume that there is an effect; a coin is biased ($p = 0.8$)
- How many times should we toss the coin to find that it is biased?
- If toss it a billion times, then surely will find that the coin is biased.
- But, we can't toss it a billion times (we are too lazy)
- What if we only toss 100 times? Can we find that the coin is biased?
  - We say the coin is biased if frequency of heads (or tails) is far from 0.5
  - If we toss only 100 times, it is possible that sometimes the frequency of heads is not 0.8, but close to 0.5 (just by randomness). In this case, we cannot say the coin is biased.
  - Let's say we will do this 100 tosses 10 times. And, suppose we know that we will find the coin is biased 7 out of 10 times.
  - Then, **the power is 70%**

Statistical Power (More formal definition)

- The power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis (Greene 2000)
- The statistical power is the ability of a test to detect an effect, if the effect actually exists (High 2000)
- Cohen (1988) says, it is the probability that it will result in the conclusion that the phenomenon exists
- If power is 100%, we will always find that the effect exists (e.g. the coin is biased)
- If power is 20%, we will find that the effect exists one out of 5 times
- Obviously, we want high power. How can we achieve high power?

Power of Association Studies

- Let's assume that SNP A is associated with the disease

$$p_A^+ - p_A^- \neq 0$$

$$p_A^+ = 0.4 \qquad p_A^- = 0.3$$

- How do we find that SNP A is associated with the disease?
  - We collect cases and controls and compute association statistic (S_A)
  - If p-value of S_A < significance threshold (0.05), we find the association

- Can we always find the association?
  - If we repeat the association study (recollect cases and controls), would we again find that p-value of S_A < significance threshold?
  - If we collect a billion cases and a billion controls for each study, we are sure that we will find the association again and again (power = 100%) because

$$\hat{p}_A^+ \simeq p_A^+ = 0.4 \qquad \hat{p}_A^- \simeq p_A^- = 0.3$$

Power of Association Studies

- What if we collect only 100 individuals for each study?

Study #	$\hat{p}_A^+$	$\hat{p}_A^-$	S_A	p-value	Is significant?
1	0.45	0.3	2.19	0.014	**Yes**
2	0.4	0.35	0.73	0.23	No
3	0.43	0.38	0.72	0.23	No
4	0.42	0.27	2.23	0.012	**Yes**
5	0.44	0.29	2.20	0.013	**Yes**

- The power is 3 / 5 = 60%
- Obviously, we want to detect the association when we collect cases and controls and compute association statistic (if it exists)
- If there are not enough individuals, we may not detect the association even if it exists

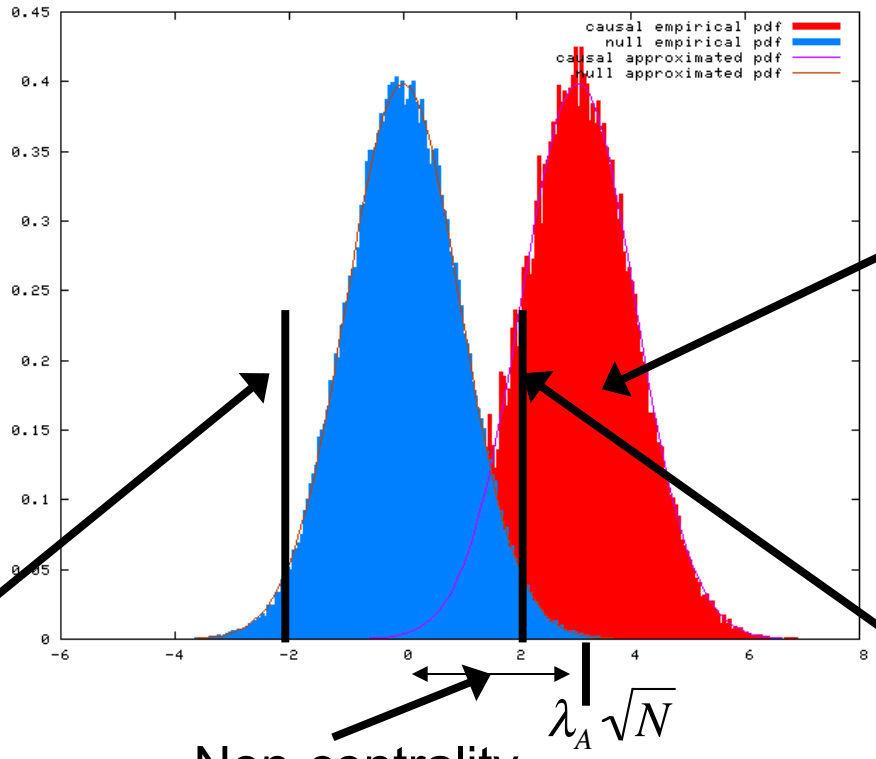
How can we compute power of association studies?

- Power of association studies is the area under the alternative distribution for Pr(X >= $\Phi^{-1}(1-\alpha/2)$) and Pr (X <= $\Phi^{-1}(\alpha/2)$) where Φ^{-1} is qnorm

$$\text{If } p_A^+ - p_A^- \neq 0 \text{ (Alt Hypothesis)}, S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2/N}\sqrt{\hat{p}_A(1-\hat{p}_A)}} \sim \mathcal{N}\left(\lambda_A\sqrt{N}, 1\right)$$

$$\text{where } \lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}, \text{ noncentrality-parameter is } \lambda_A\sqrt{N}$$

How can we compute power of association studies?



Significance threshold

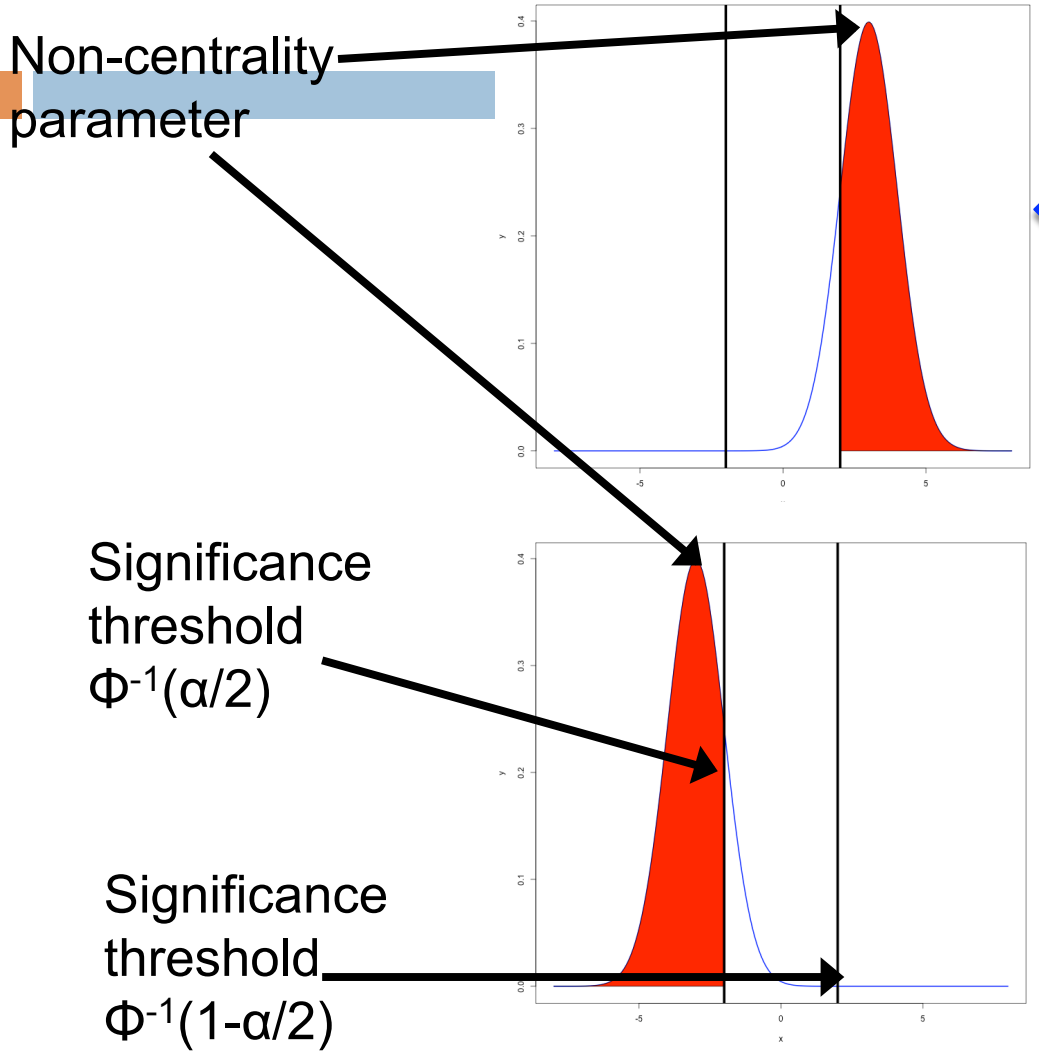
Power of association test

$\Phi^{-1}(\alpha/2)$
qnorm(0.05/2)=-1.96

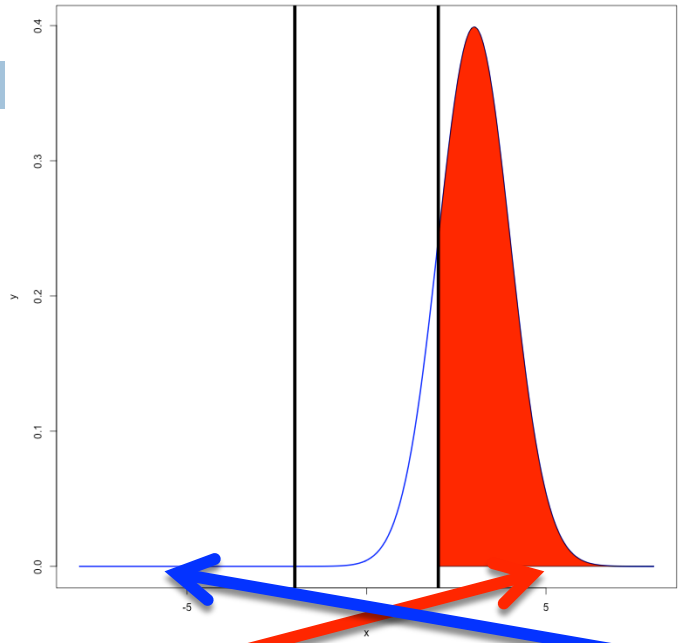
Non-centrality parameter

$\Phi^{-1}(1-\alpha/2)$
qnorm(1-0.05/2)=1.96

How can we compute power of association studies?



Power Equation



$$\text{Power} = \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N})$$

$$= \text{pnorm}(\text{qnorm}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \text{pnorm}(-\text{qnorm}(\alpha/2) + \lambda_A\sqrt{N})$$

Very Very Important Equation (Memorize for midterm)

Example of power equation

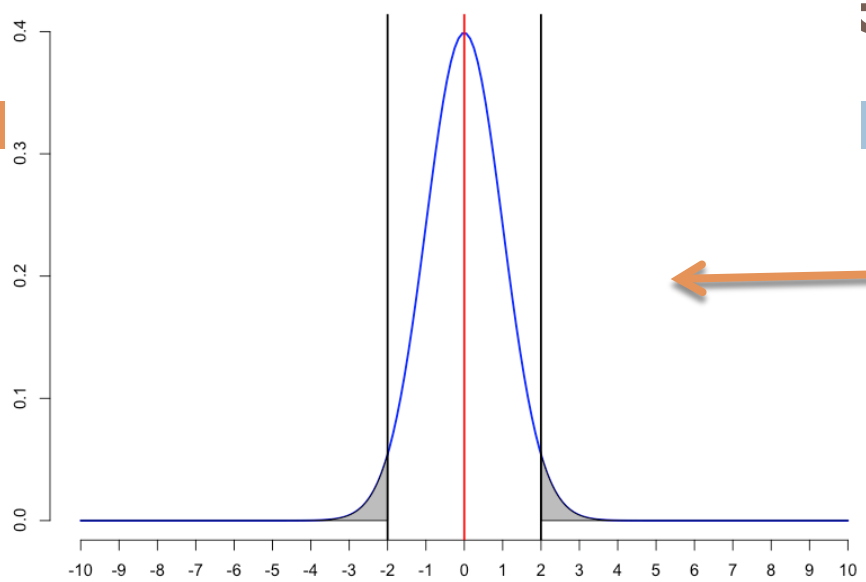
$$\text{Power} = \text{pnorm}(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})+\ \ 1-\text{pnorm}(-\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})$$

Let $\lambda_A\sqrt{N}$ be 0, then alternative distribution is centered at 0

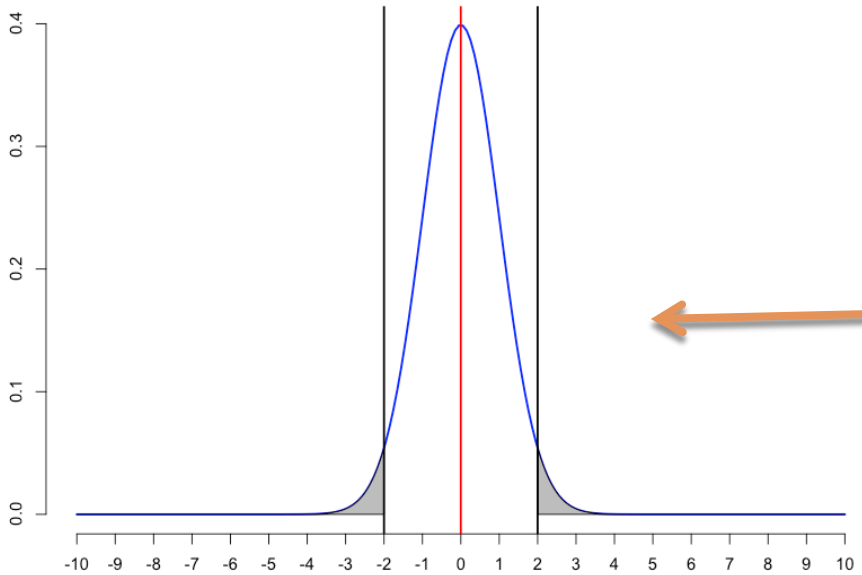
And, $(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N}) = -1.96+0 = -1.96$

So, power = $\text{pnorm}(-1.96)+1-\text{pnorm}(1.96)$=0.025+(1-0.025)=0.05

Example of power equation



What we want to compute (area under the curve from the significance threshold)



What we are computing

pnorm(-1.96)+1-pnorm(1.96)

Example of power equation

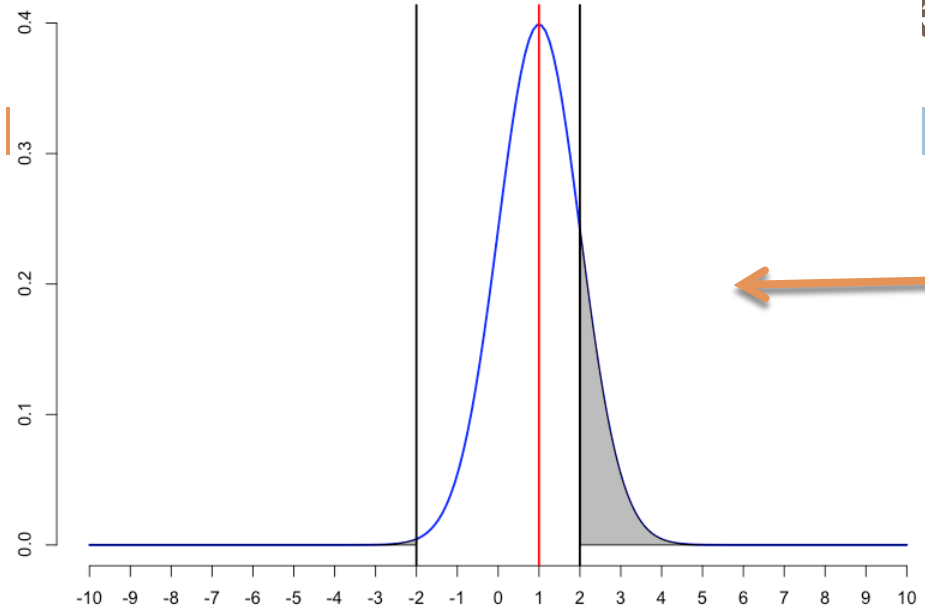
$$\text{Power} = \text{pnorm}(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})+\ \ 1-\text{pnorm}(-\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})$$

Let $\lambda_A\sqrt{N}$ be 1, then alternative distribution is centered at 1

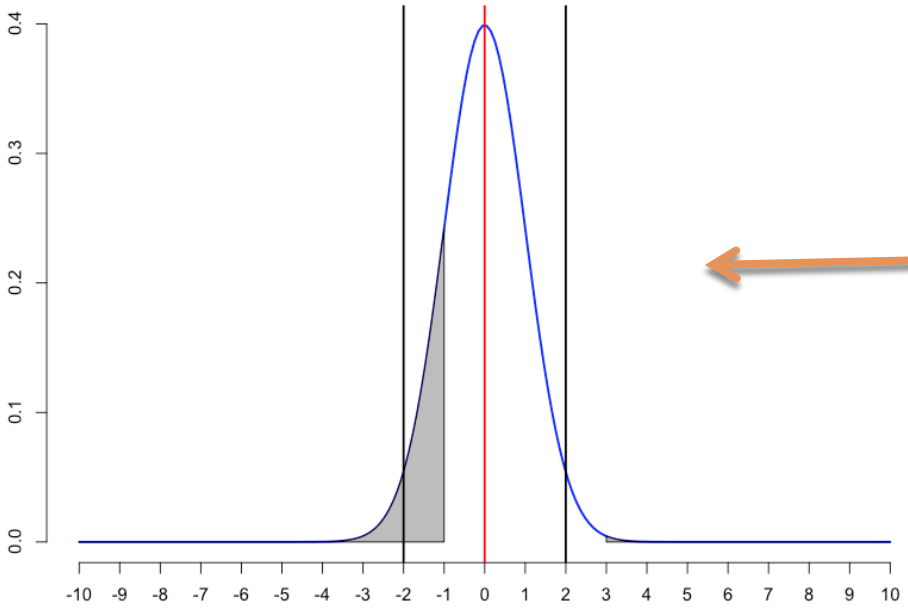
And, $(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N}) = -1.96+1 = -0.96$

So, power = $\text{pnorm}(-0.96)+1-\text{pnorm}(2.96)$=0.169+(1-0.998)=0.17

er equation



What we want to compute (area under the curve from the significance threshold)



What we are computing

pnorm(-0.96)+1-pnorm(2.96)

Example of power equation

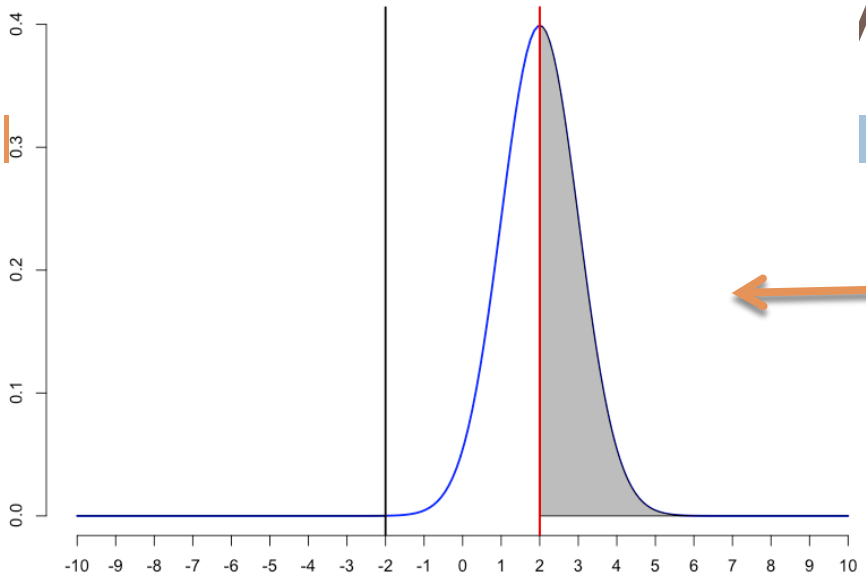
$$\text{Power} = \text{pnorm}(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})+\ \ 1-\text{pnorm}(-\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})$$

Let $\lambda_A\sqrt{N}$ be 2, then alternative distribution is centered at 2

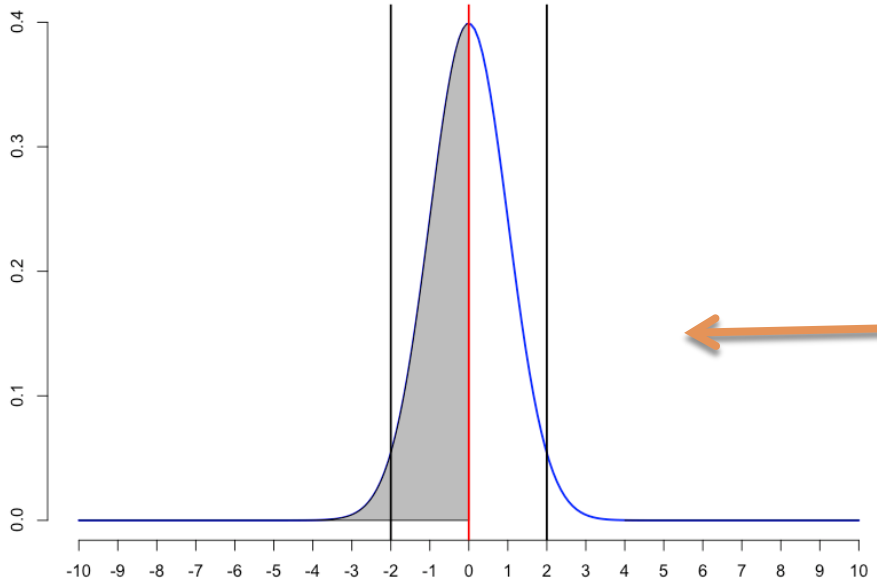
And, $(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N}) = -1.96+2 = 0.04$

So, power = $\text{pnorm}(0.04)+1-\text{pnorm}(3.96)$=0.516+(1-0.999)=0.516

ver equation



What we want to compute (area under the curve from the significance threshold)



What we are computing

pnorm(0.04)+1-pnorm(3.96)

Example of power equation

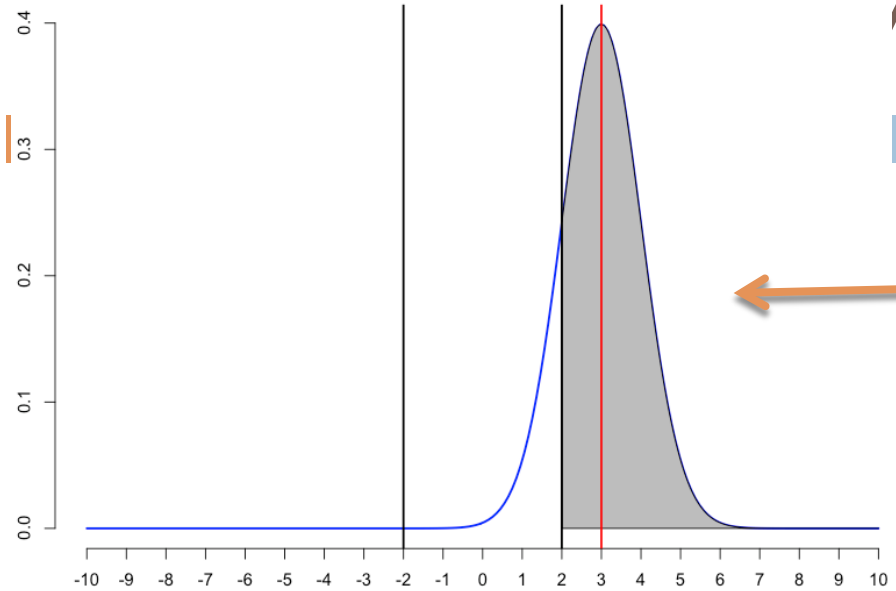
$$\text{Power} = \text{pnorm}(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})+\ \ 1-\text{pnorm}(-\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N})$$

Let $\lambda_A\sqrt{N}$ be 2, then alternative distribution is centered at 3

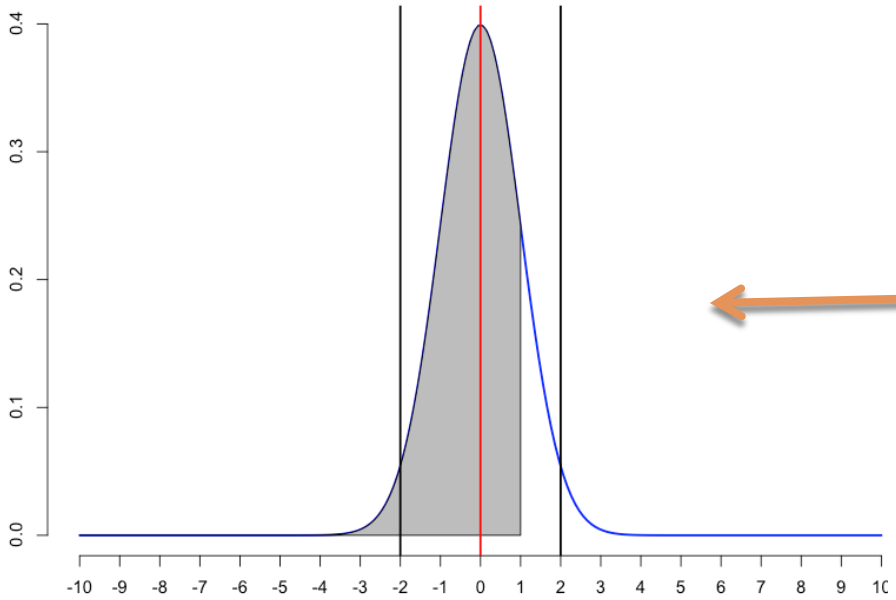
And, $(\text{qnorm}(\alpha/2)+\lambda_A\sqrt{N}) = -1.96+3 = 1.04$

So, power = $\text{pnorm}(1.04)+1-\text{pnorm}(4.96)$=0.85+(1-0.999)=0.850

er equation



What we want to compute (area under the curve from the significance threshold)



What we are computing

pnorm(1.04)+1-pnorm(4.96)

More on Power Equation

- Three things that affect the power

1. Sample size (the total number of individuals)
2. Effect size (relative risk of a SNP)
3. Significance threshold (α)

$$\text{Power} = \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) \; + \; 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N})$$

- Significance Threshold
  - Larger α, larger NCP
  - Usually given
- Effect Size
  - Larger relative risk, larger NCP
  - Usually given
- Sample Size
  - Larger sample size, larger NCP
  - More difference between λ_A and 0
  - NCP shifted further away from 0

Power Example

- Significance threshold $\alpha = 0.05$
- Assume true case frequency = 0.6, true control frequency = 0.5
- Assume we collect 100 cases and 100 controls

$$p_A^+ = 0.6 \quad p_A^- = 0.5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = 0.55 \quad N = 200$$

- Compute non-centrality parameter

$$\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{0.6 - 0.5}{\sqrt{2/200}\sqrt{0.55(1-0.55)}} = 2.01$$

- Compute power

$$\begin{aligned}
\text{Power} &= \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) \\
&= \Phi(\Phi^{-1}(0.025) + 2.01) + 1 - \Phi(-\Phi^{-1}(0.025) + 2.01) \\
&= \Phi(-1.95 + 2.01) + 1 - \Phi(1.95 + 2.01) \\
&= \Phi(0.06) + 1 - \Phi(3.96) = 0.52 + 1 - 0.9999625 = \boxed{0.52}
\end{aligned}$$

Power Example

- Significance threshold $\alpha = 0.05$
- Assume true case frequency = 0.6, true control frequency = 0.5
- Assume we collect **500** cases and **500** controls

$$p_A^+ = 0.6 \quad p_A^- = 0.5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = 0.55 \quad N = 1000$$

- Compute non-centrality parameter

$$\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{0.6 - 0.5}{\sqrt{2/1000}\sqrt{0.55(1-0.55)}} = 4.49$$

- Compute power

$$\begin{aligned}
\text{Power} &= \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) \\
&= \Phi(\Phi^{-1}(0.025) + 4.49) + 1 - \Phi(-\Phi^{-1}(0.025) + 4.49) \\
&= \Phi(-1.95 + 4.49) + 1 - \Phi(1.95 + 4.49) \\
&= \Phi(2.54) + 1 - \Phi(6.44) = 0.99 + 1 - 1 = \boxed{0.99}
\end{aligned}$$

Relative risk

Effect size of a SNP

Association Strength

- A causal SNP has a certain strength of effect on the disease.
- This effect can be parameterized by:
  γ = relative risk
- Definitions:
  p_A = allele frequency of SNP A.
  F = disease prevalence
  +/- = disease state.
- Derivation of case and control frequencies:
  $P(A)=p_A$ $p^+_A=P(A|+)$ $p^-_A=P(A|-)$ $F=P(+)$
  $P(A|+)=P(+|A)P(A)/P(+)$
  $P(+|A)= \gamma P(+|\neg A)$
  $P(+)=F=p_A P(+|A)+(1-p_A)P(+|\neg A)$
  $P(+)=F= p_A P(+|A)+(1-p_A)P(+|A)/\gamma$
  $P(+)=F=P(+|A)(p_A+(1-p_A)/\gamma)= P(+|A)(p_A(\gamma-1)+1)/\gamma$
  $P(+|A)= \gamma F/(p_A(\gamma-1)+1)$
  $P(A|+)=P(+|A)P(A)/P(+)= P(+|A)p_A/F= \gamma p_A/(p_A(\gamma-1)+1)$

$$\gamma = \frac{P(+\mid A)}{P(+\mid \neg A)}$$

Relative risk is a ratio between 1) probability of having a disease when you have a SNP and 2) probability of having a disease when you do not have a SNP

Relative Risk Examples

- Assume relative risk = 1.5
- Assume disease prevalence (F) is very small (0.001)
- Assume allele frequency (p_A) is 0.2 (sometimes called "population allele frequency")
- We can then compute true case frequency and true control frequency

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1)p_A + 1} = \frac{1.5 * 0.2}{(1.5 - 1) * 0.2 + 1} = 0.273$$

$$p_A^- = p_A = 0.2$$

- NOTE: this p_A is not the same as p_A in NCP (λ_A)

$$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1 - p_A)}}, \text{ where } p_A = \frac{p_A^+ + p_A^-}{2} \quad \left(\text{I use } p_A^{\pm} \text{ for this } p_A\right)$$

Relative Risk Examples

- Assume relative risk = 2.0
- Assume disease prevalence (F) is very small (0.001)
- Assume allele frequency (p_A) is 0.2 (sometimes called "population allele frequency")
- We can then compute true case frequency and true control frequency

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1) p_A + 1} = \frac{2 * 0.2}{(2-1) * 0.2 + 1} = 0.333$$

$$p_A^- = p_A = 0.2$$

- A larger difference between true case frequency and true control frequency
  - This example: 0.333 – 0.2 = 0.133
  - Previous example: 0.273 – 0.2 = 0.073
  - Thus, higher power for this example

HW1 Pr 1 – Part A. Calculating the NCP

Use the formulas described in Lectures 2 & 3 to compute the non-centrality parameters. Compute the non-centrality parameters for minor allele frequencies 0.05, 0.2 and 0.4, for relative risks of 1.5, 2.0 and 3.0, for total individual numbers in the cases and controls of 500 and 1000. You must compute the non-centrality parameter using R and show a transcript of your code and results. You can enter the results into these tables and include them in the homework submission.

500 individuals		Allele frequency		
		0.05	0.2	0.4
Relative risk	1.5			
	2.0			
	3.0			

Table 1: 500 Individuals

1000 individuals		Allele frequency		
		0.05	0.2	0.4
Relative risk	1.5			
	2.0			
	3.0			

Table 2: 1000 Individuals

HW1 Pr 1 – Part A. Calculating the NCP

Non-centrality parameter is $\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2p_A(1-p_A)}}\sqrt{N}$

Given p_A and relative risk (γ), we can compute

true case frequency (p_A^+) and true control frequency (p_A^-) as

$$p_A^+ = \frac{\gamma p_A}{(\gamma - 1)p_A + 1} \qquad p_A^- = p_A \qquad p_A \text{ (in } \lambda_A) = \frac{p_A^+ + p_A^-}{2}$$

You can create a R function for NCP like

```
ncp = function(gamma, pa, N) {
      pplus = (gamma*pa)/((gamma-1)*pa+1)
      pminus = pa
      ppm = (pplus+pminus)/2
      lambda = (pplus-pminus)/(sqrt(2*ppm*(1-ppm)))
      ncp = lambda*sqrt(N)
      return(ncp)
}
> ncp(1.5,0.05,500)
[1] 1.52396
```

One tip: rather than calling this function for every pair of allele frequency and relative risk, you can use "outer" function in R to compute NCP for all relative risks and frequencies. Type ?outer in R for help.

HW1 Pr 1 – Part B. Calculating the power

Now compute the power of these studies assuming a p-value threshold of 0.05. You must compute the power using R and show a transcript of your code and results. You should re-use the R code you wrote for computing non-centrality parameter in Part A. You can enter the results into these tables and include them in the homework submission,

500 individuals		Allele frequency		
		0.05	0.2	0.4
Relative risk	1.5			
	2.0			
	3.0			

Table 1: 500 Individuals

1000 individuals		Allele frequency		
		0.05	0.2	0.4
Relative risk	1.5			
	2.0			
	3.0			

Table 2: 1000 Individuals

HW1 Pr 1 – Part B. Calculating the power

Power Equation

$$= \Phi(\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A\sqrt{N})$$

$$= \text{pnorm}(\text{qnorm}(\alpha/2) + \lambda_A\sqrt{N}) + 1 - \text{pnorm}(-\text{qnorm}(\alpha/2) + \lambda_A\sqrt{N})$$

You can create R function for Power like

```
power = function(gamma, pa, N) {
    return(pnorm(qnorm(0.05/2)+ncp(gamma,pa,N))+1-pnorm(-1*qnorm(0.05/2)+ncp(gamma,pa,N)))
}
 > power(1.5,0.05,500)
 [1] 0.331664
```

One tip: rather than calling this function for every pair of allele frequency and relative risk, you can use “outer” function in R to compute NCP for all relative risks and frequencies. Type ?outer in R for help.

HW1 Pr 1 – Part C. Calculating # of individuals

Using the same relative risks and minor allele frequencies as in Part A and B, compute the number of individuals needed to achieve 80% power for each pair of relative risk and minor allele frequency. You should use the R code you wrote for Part B, and try different values of the number of individuals to achieve 80% power roughly (79% ~ 81%). You can enter the results into these tables and include them in the homework submission,

		Allele frequency		
		0.05	0.2	0.4
Relative risk	1.5			
	2.0			
	3.0			

Table 5: 80% power

- Try different values for N in the previous power function to achieve 80% power

Pr 2 – Unbalanced Cases and Controls Part A

Assume that you have N total individuals in a balanced case and control study (i.e. N/2 case individuals and N/2 control individuals). The non-centrality parameter for this study is

$$\lambda_A \sqrt{N}$$

On the other hand, if the number of cases and controls are not equal, the non-centrality parameter is different. If there are $N^+/2$ cases and $N^-/2$ controls, the non-centrality parameter is

$$\lambda_A \sqrt{\frac{2(N^+N^-)}{N^+ + N^-}}$$

Now assume you are designing a study with three times the number of cases as controls. How large does your study have to be (as a factor of N) so that you achieve the same power as a balanced study with N individuals?

Pr 2 – Unbalanced Cases and Controls Part A

In the balanced study, NCP given N total individuals is $\lambda_A \sqrt{N}$

In the unbalanced study, let N' = the total number of individuals

$\frac{N^+}{2}$ = the number of case individuals $\frac{N^-}{2}$ = the number of control individuals

N^+ = the number of case chromosomes N^- = the number of control chromosomes

$$N' = \frac{N^+}{2} + \frac{N^-}{2}, \quad N^+ + N^- = 2N', \quad \text{NCP is } \lambda_A \sqrt{\frac{2(N^+ N^-)}{N^+ + N^-}}$$

We have three times the number of cases as control, so $N^+ = 3N^-$

Re-write N^+ and N^- in terms of N'

$$\text{Eq 1) } 3N^- + N^- = 2N' \Rightarrow 4N^- = 2N' \Rightarrow N^- = (1/2)N'$$

$$\text{Eq. 2) } N^+ + (1/3)N^+ = 2N' \Rightarrow (4/3)N^+ = 2N' \Rightarrow N^+ = (3/2)N'$$

Plug N^+ and N^- into NCP of unbalanced study, and set it equal to NCP of balanced study,

$$\lambda_A \sqrt{N} = \lambda_A \sqrt{\frac{2(N^+ N^-)}{N^+ + N^-}} = \lambda_A \sqrt{\frac{2((3/2)N'(1/2)N')}{2N'}}$$

Solve for N' in terms of N

Pr 2 – Unbalanced Cases and Controls Part B

Assume that you have N+/2 cases and an unlimited number of controls. Derive what the size of the balanced study is with equivalent power. (Hint: First solve for the noncentrality parameter if you have a very large number of controls, try using 1,000,000)

In this problem, we have $N^+ / 2$ cases and an infinite number of controls

the NCP is $\lambda_A \sqrt{\frac{(2N^+N^-)}{N^+ + N^-}}$

Similar to Part A, we set NCP of balanced and unbalanced studies equal,

$$\lambda_A \sqrt{N} = \lambda_A \sqrt{\frac{2(N^+N^-)}{N^+ + N^-}}$$

$$\lambda_A \sqrt{N} = \lambda_A \sqrt{2N^+ \lim_{N^- \to \infty} \frac{N^-}{2N^+ + N^-}}$$

What happens to $\frac{N^-}{2N^+ + N^-}$ as $N^- \to \infty$?

Then, solve N in terms of $(N^+ / 2)$, the number of cases

Pr 2 – Unbalanced Cases and Controls Part C

(Grad Students ONLY)

Derive the non-centrality parameter for unbalanced cases and controls above. Describe the precise approximation assumption you need to make.

$$\hat{p}_A^+ \sim N(p_A^+, p_A^+(1-p_A^+)/N^+)$$

$$\hat{p}_A^- \sim N(p_A^-, p_A^-(1-p_A^-)/N^-)$$

Taking the difference,

$$\hat{p}_A^+ - \hat{p}_A^- \sim \mathcal{N}\left(p_A^+ - p_A^-, \frac{N^- p_A^+(1-p_A^+) + N^+ p_A^-(1-p_A^-)}{N^+ N^-}\right)$$

We use the following approximation

$$N^- p_A^+(1-p_A^+) + N^+ p_A^-(1-p_A^-) \approx (N^- + N^+)(p_A(1-p_A))$$

$$\hat{p}_A^+ - \hat{p}_A^- \sim N\left(p_A^+ - p_A^-, \frac{(N^- + N^+)\, p_A(1-p_A)}{N^+ N^-}\right)$$

Divide the equation by the square root of variance term so that variance is 1

Then, after doing some algebraic manipulation, you can show that

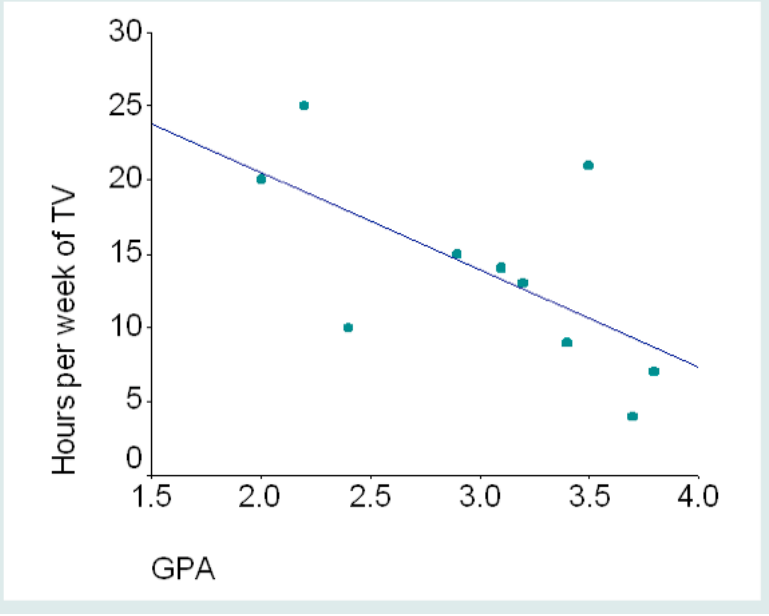
NCP is $\lambda_A \sqrt{\frac{2(N^+ N^-)}{N^+ + N^-}}$

Correlation

- What is a correlation (in general)?
  - A correlation is a single number that describes the degree of relationship between two variables
  - Ranges from –1.00 to +1.00 (often denoted as *r*)
- Example (GPA vs. TV in hours per week) from (http://www.nvcc.edu/home/elanthier/methods/correlation.htm)

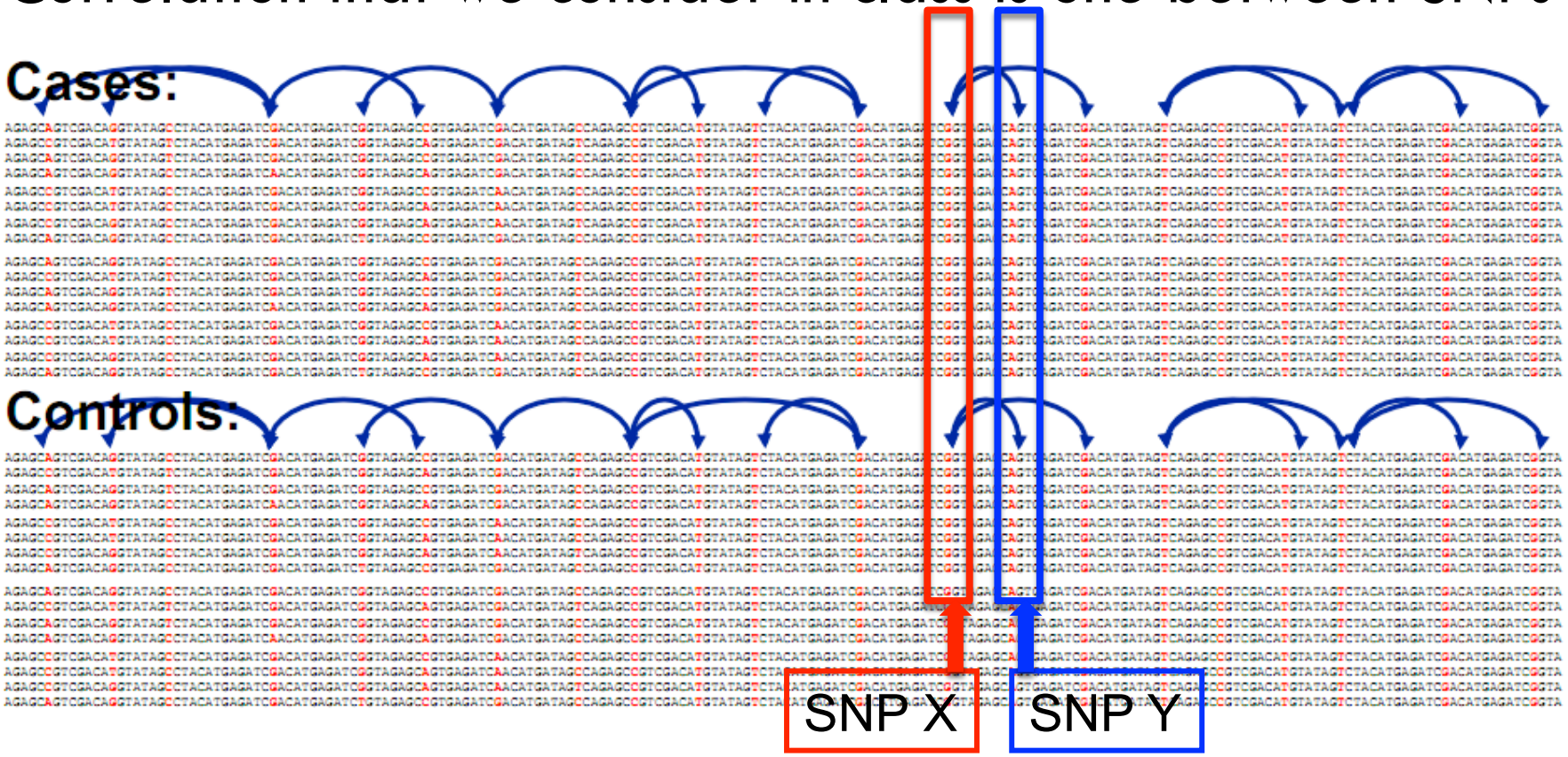
Participant	GPA	TV in hours per week
#1	3.1	14
#2	2.4	10
#3	2.0	20
#4	3.8	7
#5	2.2	25
#6	3.4	9
#7	2.9	15
#8	3.2	13
#9	3.7	4
#10	3.5	21

In this sample, the correlation is -.63.



Correlation (Linkage Disequilibrium)

- Correlation that we consider in class is one between SNPs



Correlation (Linkage Disequilibrium)

Ind	SNP X	SNP Y
1	A	C
2	A	C
3	A	C
4	T	G
5	T	G
6	T	G
7	T	G
8	T	G
9	T	G
10	T	G

- Perfect correlation
- If you have A allele in SNP X, you always have C allele in SNP Y
- If you have T allele in SNP X, you always have G allele in SNP Y
- SNP X and SNP Y have $r = 1$ and they are in *linkage disequilibrium*
- Implication: we do not need to collect information about SNP Y if we collect SNP X

Correlation (Linkage Disequilibrium)

Ind	SNP X	SNP Y
1	A	C
2	T	G
3	A	C
4	T	G
5	T	G
6	T	C
7	A	G
8	T	G
9	T	G
10	T	G

- $r = 0.52$
- Assume SNP Y is causal, but collect SNP X (why not collect Y? we'll discuss later)
- Suppose we collect SNP Y with 1000 individuals and we know we achieve 90% power (the probability of detecting that SNP Y is associated with a disease)
- What would be the power of detecting association of Y if we collect SNP X?
- Intuitively, the closer X is to Y (higher *r*), the higher power
- The more X is different from Y (lower *r*), the lower power

Indirect Association

- Assume we have two SNPs, A and B
- B is the causal SNP (two alleles are B and b)
- However, we collect A (two alleles are A and a)

- We want to relate

$$\lambda_A = \frac{(p_A^+ - p_A^-)}{\sqrt{2p_A(1-p_A)}} \qquad S_A \sim N\left(\lambda_A\sqrt{N},1\right)$$

- to

$$\lambda_B = \frac{(p_B^+ - p_B^-)}{\sqrt{2p_B(1-p_B)}} \qquad S_B \sim N\left(\lambda_B\sqrt{N},1\right)$$

Indirect Association (derivation)

- Most difficult problem (in terms of length) in the midterm
- One key assumption: conditional probability distributions are equal in cases and controls

$$p^{+}_{A|B} = p^{-}_{A|B} = p_{A|B}$$

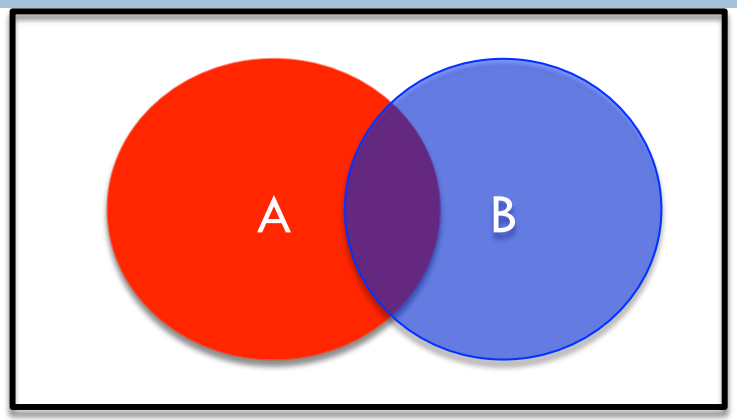
1. Let's write the true case frequency at SNP A in terms of joint probabilities of SNPs A and B

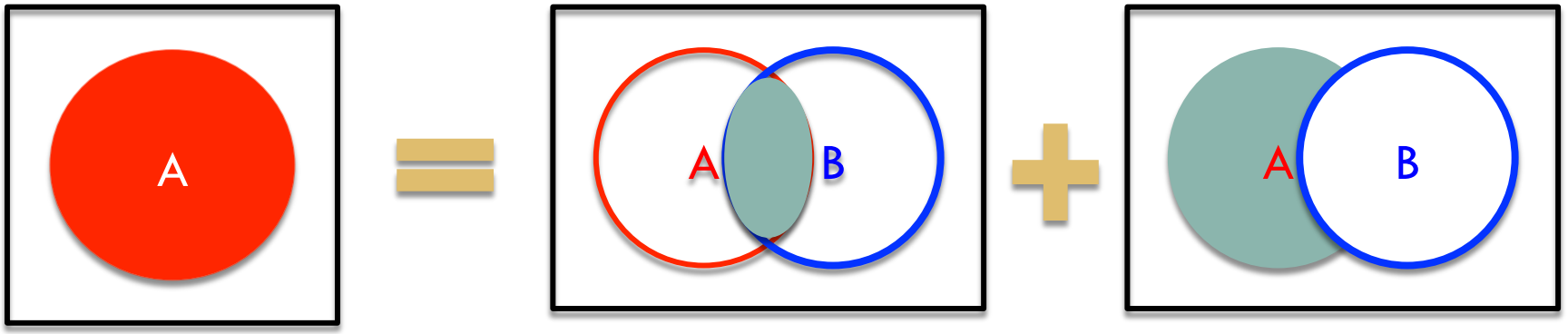
$$p^{+}_{A} = p^{+}_{AB} + p^{+}_{Ab}$$

Let's understand this equation in terms of Venn diagram

Indirect Association (derivation)

$$p_A^+ = p_{AB}^+ + p_{Ab}^+$$





$$p(A) = p(A \cap B) + p(A \cap \neg B)$$

Indirect Association (derivation)

2. Use conditional probability

$$p_{A|B} = \frac{p_{AB}}{p_B} \Leftrightarrow p_{AB} = p_B p_{A|B}$$

$$p_{Ab} = p_b p_{A|b} = (1 - p_B) p_{A|b} \quad \text{because } p_b = 1 - p_B$$

3. Rewrite p^+_{AB} and p^-_{AB}

$$p_A^+ = p_{AB}^+ + p_{Ab}^+$$

$$p_A^+ = p_B^+ p_{A|B} + (1 - p_B^+) p_{A|b} \qquad \text{Remember: } p_{A|B}^+ = p_{A|B}^- = p_{A|B}$$

$$p_A^- = p_B^- p_{A|B} + (1 - p_B^-) p_{A|b}$$

Indirect Association (derivation)

4. Take a difference between p^+_A and p^-_A

$$p^+_A = p^+_B p_{A|B} + (1 - p^+_B) p_{A|b} \quad - \quad p^-_A = p^-_B p_{A|B} + (1 - p^-_B) p_{A|b}$$

$$p^+_A - p^-_A = p^+_B p_{A|B} + (1 - p^+_B) p_{A|b} - p^-_B p_{A|B} - (1 - p^-_B) p_{A|b}$$

$$= p^+_B p_{A|B} + p_{A|b} - p^+_B p_{A|b} - p^-_B p_{A|B} - p_{A|b} + p^-_B p_{A|b} \quad \text{(expand all terms)}$$

$$= p^+_B p_{A|B} - p^-_B p_{A|B} - p^+_B p_{A|b} + p^-_B p_{A|b} \quad (p_{A|b} \text{ canceled})$$

$$= p_{A|B}(p^+_B - p^-_B) - p_{A|b}(p^+_B - p^-_B) \quad \text{(arrage terms)}$$

$$= (p^+_B - p^-_B)(p_{A|B} - p_{A|b}) \quad \text{(arrage terms)}$$

Indirect Association (derivation)

5. Substitute $p^+_A - p^-_A$ into λ_A

$$\lambda_A = \frac{p^+_A - p^-_A}{\sqrt{2p_A(1-p_A)}} \quad \text{and} \quad p^+_A - p^-_A = (p^+_B - p^-_B)(p_{A|B} - p_{A|b})$$

$$\lambda_A = \frac{(p^+_B - p^-_B)(p_{A|B} - p_{A|b})}{\sqrt{2p_A(1-p_A)}}$$

$$= \frac{(p^+_B - p^-_B)(p_{A|B} - p_{A|b})}{\sqrt{2p_A(1-p_A)}} \frac{\sqrt{2p_B(1-p_B)}}{\sqrt{2p_B(1-p_B)}} \qquad \left(\text{multiply by } \frac{\sqrt{2p_B(1-p_B)}}{\sqrt{2p_B(1-p_B)}} = 1 \right)$$

$$= \frac{(p^+_B - p^-_B)}{\sqrt{2p_B(1-p_B)}} \frac{(p_{A|B} - p_{A|b})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}} \quad \text{(arrange terms)}$$

$$= \lambda_B \frac{(p_{A|B} - p_{A|b})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}} \qquad \left(\lambda_B = \frac{(p^+_B - p^-_B)}{\sqrt{2p_B(1-p_B)}} \right)$$

Indirect Association (derivation)

6. Conditional probability again

$$p_{A|B} = \frac{p_{AB}}{p_B} \text{ and } p_{A|b} = \frac{p_{Ab}}{p_b} = \frac{p_{Ab}}{1-p_B} \quad \text{because } p_b = 1 - p_B$$

Then

$$\lambda_A = \lambda_B \frac{(p_{A|B} - p_{A|b})\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

$$\lambda_A = \lambda_B \frac{\left(\frac{p_{AB}}{p_B} - \frac{p_{Ab}}{1-p_B}\right)\sqrt{2p_B(1-p_B)}}{\sqrt{2p_A(1-p_A)}}$$

Indirect Association (derivation)

$$\lambda_A = \lambda_B \frac{\left(\frac{p_{AB}}{p_B} - \frac{p_{Ab}}{1-p_B}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}} = \lambda_B \frac{\left(\frac{p_{AB}(1-p_B)}{p_B(1-p_B)} - \frac{p_{Ab}p_B}{(1-p_B)p_B}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

$$= \lambda_B \frac{\left(\frac{p_{AB}(1-p_B) - p_{Ab}p_B}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}} = \lambda_B \frac{\left(\frac{p_{AB} - p_{AB}p_B - p_{Ab}p_B}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

$$= \lambda_B \frac{\left(\frac{p_{AB} - p_B(p_{AB} + p_{Ab})}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}} = \lambda_B \frac{\left(\frac{p_{AB} - p_B p_A}{p_B(1-p_B)}\right)\sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}}$$

Remember $p_A = p_{AB} + p_{Ab}$

Indirect Association (derivation)… Finally

$$\lambda_A = \lambda_B \frac{\left(\frac{p_{AB} - p_B p_A}{p_B(1-p_B)} \right) \sqrt{p_B(1-p_B)}}{\sqrt{p_A(1-p_A)}} = \lambda_B \frac{\left(\frac{p_{AB} - p_B p_A}{\sqrt{p_B(1-p_B)}} \right)}{\sqrt{p_A(1-p_A)}}$$

$$= \lambda_B \frac{p_{AB} - p_B p_A}{\sqrt{p_B(1-p_B)}\sqrt{p_A(1-p_A)}} = \lambda_B \sqrt{r^2}$$

Indirect Association (example)

- Significance threshold $\alpha = 0.05$
- Causal SNP A: true case/control probabilities are 0.6/0.5
- Collect SNP B and r^2_{AB}=0.8
- Collect 400 case and 400 control individuals
- First calculate NCP at the causal SNP (A)

$$p_A^+ = .6 \quad p_A^- = .5 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .55 \quad N = 800$$

$$\lambda_A \sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.6-.5}{\sqrt{2/800}\sqrt{.55(1-.55)}} = 4.02$$

- Next calculate NCP at the collected SNP (B)

$$\lambda_B \sqrt{N} = \lambda_A \sqrt{N} \sqrt{r^2_{AB}} = 4.02 * \sqrt{.8} = 3.56$$

Indirect Association (example)

- Compute power using NCP of SNP B

$$\text{Power} = \Phi(\Phi^{-1}(\alpha/2) + \lambda_B\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_B\sqrt{N})$$

$$= \Phi(\Phi^{-1}(0.025) + 3.56) + 1 - \Phi(-\Phi^{-1}(0.025) + 3.56)$$

$$= .95$$

Multiple testing with SNPs

- Each coin corresponds to each SNP
- We do not know which SNP causes a disease (we are tying to find which SNP causes a disease)
- There are more than a million SNPs
- If we look at only one SNP (that does not cause a disease), the probability that we find the SNP is associated with a disease is 0.05 (the same as one fair coin)
- If we look at a million SNPs (that do not cause a disease), the probability that we find **any SNP out of a million** is associated with a disease is much greater than 0.05 (similar to 100 coins)
- So, without "multiple hypothesis correction," we would have a lot of false positives

Multiple Hypothesis Testing "Correction"

- We want to find a new significance threshold (α_s) for each SNP such that the overall false positive rate (considering M SNPs) is $\alpha = 0.05$
- Two correction methods: Sidak and Bonferroni corrections
  - Sidak correction

$$\alpha_s = 1 - \sqrt[M]{1-\alpha}$$

  - Bonferroni correction

$$\alpha_s = \frac{\alpha}{M}$$

  - Both have similar values if M is large

Multi-SNP Association Example

- Collect data at 5 SNPs
- Significance Threshold α=0.05
- Sample: 100 Cases and 100 Controls
- Total of 200 Case Chromosomes and 200 Control Chromosomes

$\hat{p}_1^+ = \frac{120}{200} = .6$ $\hat{p}_2^+ = \frac{80}{200} = .4$ $\hat{p}_3^+ = \frac{60}{200} = .3$ $\hat{p}_4^+ = \frac{100}{200} = .5$ $\hat{p}_5^+ = \frac{120}{200} = .6$

$\hat{p}_1^- = \frac{100}{200} = .5$ $\hat{p}_2^- = \frac{75}{200} = .375$ $\hat{p}_3^- = \frac{65}{200} = .325$ $\hat{p}_4^- = \frac{95}{200} = .475$ $\hat{p}_5^- = \frac{125}{200} = .625$

$\hat{p}_1 = .55$ $\hat{p}_2 = .3825$ $\hat{p}_3 = .3125$ $\hat{p}_4 = .4875$ $\hat{p}_5 = .6125$

$$S_1 = \frac{.6 - .5}{\sqrt{2/200}\sqrt{.55(1-.55)}} = 2.01 \qquad S_2 = \frac{.4 - .375}{\sqrt{2/200}\sqrt{.3825(1-.3825)}} = .514 \qquad S_3 = \frac{.3 - .325}{\sqrt{2/200}\sqrt{.3125(1-.3125)}} = -.54$$

$$S_4 = \frac{.5 - .475}{\sqrt{2/200}\sqrt{.4875(1-.4875)}} = .500 \qquad S_5 = \frac{.6 - .625}{\sqrt{2/200}\sqrt{.6125(1-.6125)}} = -0.513$$

$S_1=S_{max}=2.01$ (Is this significant?)
Per-marker threshold $\alpha_s=\alpha/5=0.01$ (Bonferroni)
$-\Phi^{-1}(0.01/2)=2.57$
Association is not significant

When testing multiple SNPs, remember the multiple hypothesis testing

Multiple Hypothesis Testing Correction

- Bonferroni correction assumes that all tests are independent: all SNPs or all coins are independent
- As you see in the indirect association, SNPs are not independent (there is a correlation)
- Bonferroni is conservative when SNPs are not independent
  - α_s from Bonferroni (α / M) < true α_s that gives overall α
  - In other words, if we use α_s from Bonferroni on correlated SNPs, the overall false positive rate would be less than α
  - Isn't it a good thing because we have fewer false positives?
  - It's good in terms of false positives but not good in terms of power
  - Remember as the significance threshold decreases, power decreases
  - We need more number of individuals to detect that a SNP is associated with a disease if it indeed causes a disease

Multi-SNP Power analysis

- Until now, we considered power of one SNP
- We know now how to find the significance threshold when we test multiple SNPs
- We can then compute power of our association study that involves multiple SNPs
- In a Multi-SNP power problem, we are given
  - The number of SNPs (M)
  - Minor allele frequency of each SNP
  - Relative risk of a causal SNP
  - The number of cases and controls
  - The overall significance threshold
- Let's solve the problem with an example

Multi-SNP power without Tag SNPs

- Assume that we have 5 independent SNPs, 3 have minor allele frequency of .4 and 2 have a minor allele frequency of .2. Assume that the relative risk of one of them is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?
- 4 steps for solving this problem
  - 1. Compute p^+_A, p^-_A and p_A for each MAF using relative risk and MAF
  - 2. Compute NCP for each MAF using p^+_A, p^-_A, p_A and N
  - 3. Compute power for each MAF using NCP and α (don't forget Bonferroni correction!)
  - 4. Average power to compute total power using power of each MAF

Step 1: Compute p^+_A, p^-_A and p_A for each MAF using relative risk and MAF

MultiSNP Power

- If a SNP with minor allele frequency of .4 is causal, then

$$p_A^+ = \frac{\gamma p}{(\gamma - 1)p + 1} = \frac{2 * .4}{(2-1).4 + 1} = .57 \quad p_A^- = p = .4 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .485$$

- If a SNP with minor allele frequence of .2 is causal, then

$$p_A^+ = \frac{\gamma p}{(\gamma - 1)p + 1} = \frac{2 * .2}{(2-1).2 + 1} = .33 \quad p_A^- = p = .2 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .266$$

Step 2: Compute NCP for each MAF using p^{+}_{A}, p^{-}_{A}, p_A and N

MultiSNP Power

- If a SNP with minor allele frequency of .4 is causal, then

$$\lambda_{p=.4}\sqrt{N} = \frac{p_A^{+} - p_A^{-}}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.57 - .4}{\sqrt{2/200}\sqrt{.485(1-.485)}} = 3.4$$

- If a SNP with minor allele frequence of .2 is causal, then

$$\lambda_{p=.2}\sqrt{N} = \frac{p_A^{+} - p_A^{-}}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.33 - .2}{\sqrt{2/200}\sqrt{.266(1-.266)}} = 2.9$$

Step 3: Compute Power for each MAF using NCP and α (don't forget Bonferroni correction!)

MultiSNP Power

- If α=0.05, then the per-marker threshold using the Bonferroni correction, $\alpha_s = \alpha/5 = 0.01$.
- The power at a SNP with minor allele frequency 0.4 is

$$\begin{aligned}\text{power} &= \Phi(\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) \\ &= \Phi(\Phi^{-1}(0.005) + 3.4) + 1 - \Phi(-\Phi^{-1}(0.005) + 3.4) \\ &= .795\end{aligned}$$

- At a SNP with minor allele frequency 0.2

$$\begin{aligned}\text{power} &= \Phi(\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha_s/2) + \lambda\sqrt{N}) \\ &= \Phi(\Phi^{-1}(0.005) + 2.9) + 1 - \Phi(-\Phi^{-1}(0.005) + 2.9) \\ &= .627\end{aligned}$$

Step 4: Average Power to compute total power using power of each MAF

MultiSNP Power

- Since there are 3 SNPs with minor allele frequence 0.4 and 2 SNPs with minor allele frequency 0.2, the total power is

$$\text{total power} = \frac{3 * .795 + 2 * .627}{5} = .728$$

Tag SNP Selection

- HapMap found 1 ~ 2 million SNPs in humans
- Turns out that many of them are correlated
- It means that we do not need to collect 1 ~ 2 million SNPs when we do association study
  - Maybe we only need 0.5 million SNPs, which is cheaper than collecting 1 or 2 million SNPs
- Tag SNPs are ones that we actually collect in the association study
- Since we are not collecting all SNPs, tag SNPs should cover as many SNPs as possible

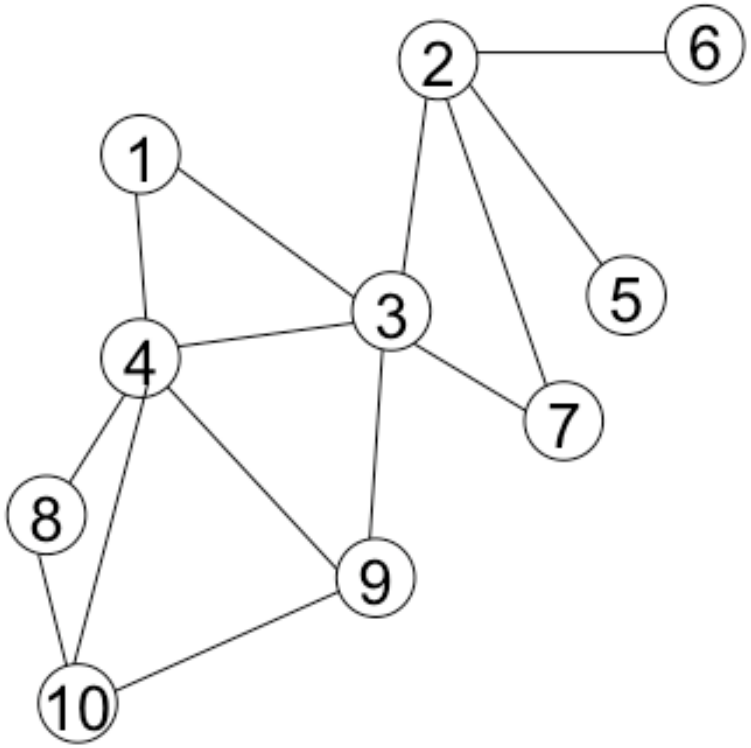
Tag SNP Selection

- We are given M SNPs, and correlation between every pair of SNPs
- We want to choose a minimum set of SNPs (called “Tag SNPs”) that covers every SNP; each SNP is either Tag SNP or has correlation value higher than some threshold with Tag SNP
- Greedy algorithm chooses SNP that is correlated with the most remaining untagged SNPs as Tag SNP until every SNP is either Tag SNP or correlated with Tag SNP
- Greedy algorithm not optimal, but good performance

Greedy Tag SNP Selection

Nodes are SNPs
Edges denote $r^2>.8$

Out Degree Counts
1: 2
2: 4
3: 5 (highest)
4: 5 (highest)
5: 1
6: 1
7: 2
8: 2
9: 3
10: 3

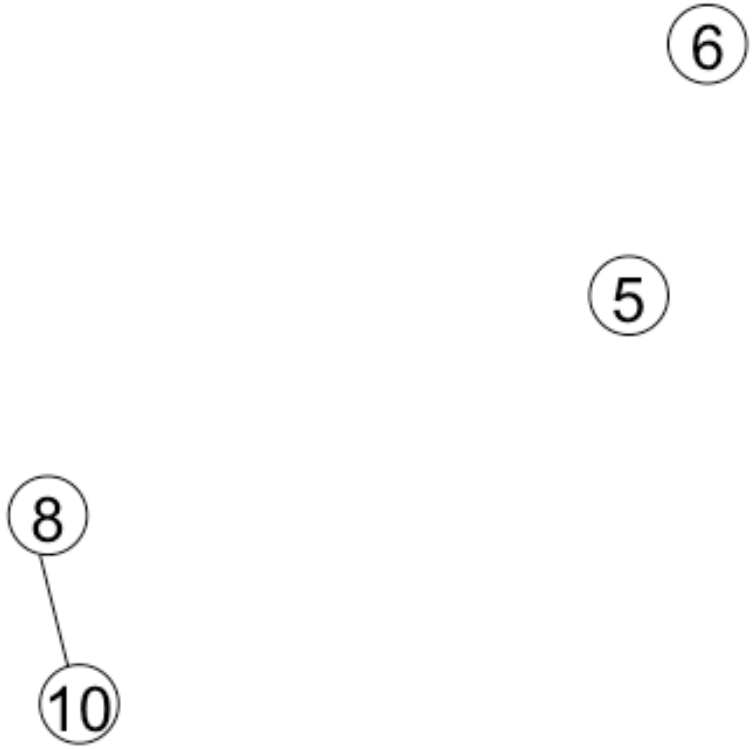


Tags 3

Greedy Tag SNP Selection

Nodes are SNPs
Edges denote $r^2>.8$

Out Degree Counts
5: 0
6: 0
8: 1 (highest)
10: 1



Tags 3,8

Greedy Tag SNP Selection

Nodes are SNPs
Edges denote $r^2>.8$

Out Degree Counts
5: 0 (highest)
6: 0

Tags 3,5,8

Greedy Tag SNP Selection

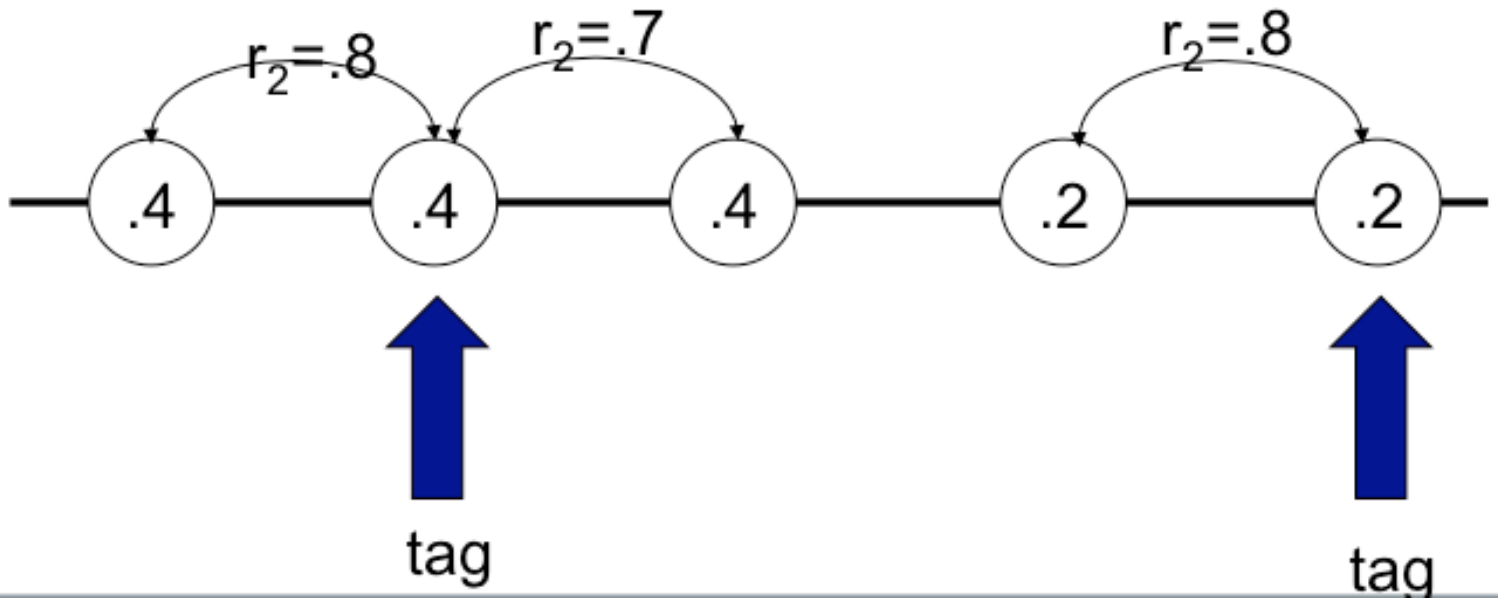
Nodes are SNPs
Edges denote r^2>.8

Out Degree Counts
6: 0 (highest)

Tags 3,5,6,8

MultiSNP Power with Tags

- Assume you have 5 SNPs, 2 of them are tags. Assume that the relative risk of one of them is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With α=0.05, what is the power of this association study?



Multi-SNP power with Tag SNPs

- 4 steps for solving this problem
  - 1. Compute p^{+}_{A}, p^{-}_{A} and p_A for each MAF using relative risk and MAF
  - 2. Compute NCP for each Tag SNP using p^{+}_{A}, p^{-}_{A}, p_A,N, and NCP for non-tagged SNP using NCP of Tag SNP and its correlation to Tag SNP
  - 3. Compute Power for each SNP using NCP and α (don't forget Bonferroni correction & the number of tag SNPs!)
  - 4. Average Power to compute total power using power of each SNP

Step 1: Compute p^+_A, p^-_A and p_A for each MAF using relative risk and MAF

MultiSNP Power

- If a SNP with minor allele frequency of .4 is causal, then

$$p_A^+ = \frac{\gamma p}{(\gamma - 1)p + 1} = \frac{2 * .4}{(2-1).4 + 1} = .57 \quad p_A^- = p = .4 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .485$$

- If a SNP with minor allele frequence of .2 is causal, then

$$p_A^+ = \frac{\gamma p}{(\gamma - 1)p + 1} = \frac{2 * .2}{(2-1).2 + 1} = .33 \quad p_A^- = p = .2 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .266$$

Step 2: Compute NCP for each Tag SNP using p^{+}_{A} , p^{-}_{A}, p_A,N, and NCP for non-tagged SNP using NCP of Tag SNP and its correlation to Tag SNP

MultiSNP Power

- If a SNP with minor allele frequency of .4 is causal, then

$$\lambda_{p=.4}\sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.57-.4}{\sqrt{2/200}\sqrt{.485(1-.485)}} = 3.4$$

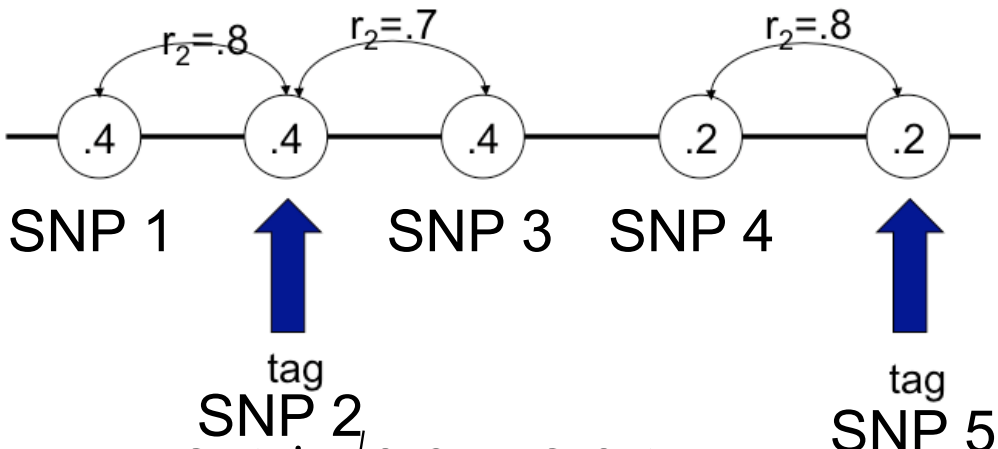
- If a SNP with minor allele frequence of .2 is causal, then

$$\lambda_{p=.2}\sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.33-.2}{\sqrt{2/200}\sqrt{.266(1-.266)}} = 2.9$$

Multi-SNP power with Tag SNPs

Jae Hoon Sul

- Step 2: compute NCP for non-tagged SNP using NCP of tag SNP and its correlation to Tag SNP



- NCP at SNP 1 = 3.4 * √0.8 = 3.04
- NCP at SNP 2 = 3.4 * √1 = 3.4 (Tag SNP)
- NCP at SNP 3 = 3.4 * √0.7 = 2.84
- NCP at SNP 4 = 2.9 * √0.8 = 2.59
- NCP at SNP 5 = 2.9 * √01 = 2.9 (Tag SNP)

Multi-SNP power with Tag SNPs

- Step 3: Compute Power for each SNP using NCP and α (don't forget Bonferroni correction & the number of tag SNPs)
  - Since there are 2 tag SNPs, $\alpha_s = \alpha / 2 = 0.05/2 = 0.025$

$$\text{power at SNP } 1 = \Phi(\Phi^{-1}(0.0125)+3.04)+1-\Phi(-\Phi^{-1}(0.0125)+3.04) = .787$$

$$\text{power at SNP } 2 = \Phi(\Phi^{-1}(0.0125)+3.4)+1-\Phi(-\Phi^{-1}(0.0125)+3.4) = .877$$

$$\text{power at SNP } 3 = \Phi(\Phi^{-1}(0.0125)+2.84)+1-\Phi(-\Phi^{-1}(0.0125)+2.84) = .725$$

$$\text{power at SNP } 4 = \Phi(\Phi^{-1}(0.0125)+2.59)+1-\Phi(-\Phi^{-1}(0.0125)+2.59) = .636$$

$$\text{power at SNP } 5 = \Phi(\Phi^{-1}(0.0125)+2.9)+1-\Phi(-\Phi^{-1}(0.0125)+2.9) = .745$$

- Step 4: Average Power to compute total power using power of each SNP

  Total Power = (0.787+0.877+0.725+0.636+0.745)/5 = **0.754**

HW2 Pr 1 – Multiple Hypothesis Testing

In class, we talked about two methods to correct for multiple hypothesis testing, Sidak and Bonferroni. Consider a multi-SNP association study where one is interested in looking for *any* SNP that is associated with a disease phenotype with a probability of 0.05 or 0.01. Compute the thresholds for association at each individual SNP if the researcher decides to consider 2, 5, 10, 100 and 1000 SNPs using both Sidak and Bonferroni corrections. Assume that the SNPs are independent.

		Significant Threshold			
		Sidak		Bonferroni	
		0.05	0.01	0.05	0.01
	2				
	5				
Number of SNPs	10				
	100				
	1000				

HW2 Pr 1 – Multiple Hypothesis Testing

Sidak Correction : $\alpha_s = 1 - \sqrt[M]{1-\alpha}$

Bonferroni Correction : $\alpha_s = \frac{\alpha}{M}$

You can create R function for Sidak and Bonferroni like

```
sidak = function(alpha,M) {
return(1-(1-alpha)^(1/M))
}

bonf = function(alpha,M) {
return(alpha/M)
}
```

Using outer function in R,

```
alpha = c(0.05,0.01)
M = c(2,5,10,100,1000)
outer(alpha,M,sidak)
outer(alpha,M,bonf)
```

HW2 Pr 2 – Tag SNP Selection Problem

We are given the following matrix of correlations, r, between 10 SNPs.

	1	2	3	4	5	6	7	8	9	10
1	1	0.9	0.85	0.5	0.4	0.2	0.2	0.15	0.15	0.1
2		1	0.95	0.5	0.8	0.2	0.2	0.15	0.15	0.1
3			1	0.65	0.9	0.7	0.5	0.5	0.3	0.2
4				1	0.85	0.5	0.85	0.6	0.7	0.7
5					1	0.75	0.6	0.75	0.6	0.5
6						1	0.6	0.75	0.4	0.3
7							1	0.8	0.85	0.8
8								1	0.6	0.5
9									1	0.5
10										1

2.1 Computing Power

Assume that we collect all 10 SNPs and the minor allele frequency (MAF) of SNPs 1 to 5 is 0.3 and MAF of SNPs 6 to 10 is 0.15. Assume that the relative risk of one of them is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?

Remember 4 steps !

HW2 Pr 2.1 – Computing Power

Step 1. Compute p^+_A, p^-_A and p_A for each MAF using relative risk and MAF

$$p_A^+ = \frac{\gamma p}{(\gamma-1)p+1} = \frac{2*.3}{(2-1).3+1} = .46 \quad p_A^- = p = .3 \quad p_A = \frac{p_A^+ + p_A^-}{2} = .38$$

```
pplus = function(gamma,p) {
return((gamma*p)/(((gamma-1)*p+1)))
}
```

Step 2. Compute NCP for each MAF using p^+_A, p^-_A, p_A and N

$$\lambda_{p=.3}\sqrt{N} = \frac{p_A^+ - p_A^-}{\sqrt{2/N}\sqrt{p_A(1-p_A)}} = \frac{.46-.3}{\sqrt{2/200}\sqrt{.38(1-.38)}} = 3.32$$

```
ncp = function(gamma,p,N) {
pp = pplus(gamma,p)
pa = (pp+p)/2
return((pp-p)/(sqrt(2/N)*sqrt(pa*(1-pa))))
}
```

HW2 Pr 2.1 – Computing Power

Step 3. Compute Power for each MAF using NCP and α (don't forget Bonferroni correction!)

If α =0.05, then the per-marker threshold using the Bonferroni correction, α_s= α/10=0.005

The power at a SNP with minor allele frequency 0.3 is

$$\text{power} = \Phi(\Phi^{-1}(\alpha_s/2)+\lambda\sqrt{N})+1-\Phi(-\Phi^{-1}(\alpha_s/2)+\lambda\sqrt{N})$$

$$= \Phi(\Phi^{-1}(0.0025)+3.32)+1-\Phi(-\Phi^{-1}(0.0025)+3.32) = .69$$

```
power = function(gamma, pa, N, alpha, M) {
    return(pnorm(qnorm(alpha/M/2)+ncp(gamma,pa,N))+1-pnorm(-1*qnorm(alpha/M/2)+ncp(gamma,pa,N)))
}
```

Step 4. Average Power to compute total power using power of each MAF

$$\text{total power} = \frac{5*.69+5*?}{10} = ?$$

```
totalpower = function(M1,p1,M2,p2,gamma,N,alpha) {
 M = M1+M2
 firstpower = power(gamma,p1,N,alpha,M)
 secondpower = power(gamma,p2,N,alpha,M)
 return((M1*firstpower+M2*secondpower)/M)
 }
```

HW2 Pr 2.2 – Greedy algorithm

2.2.1 Finding Tag SNPs

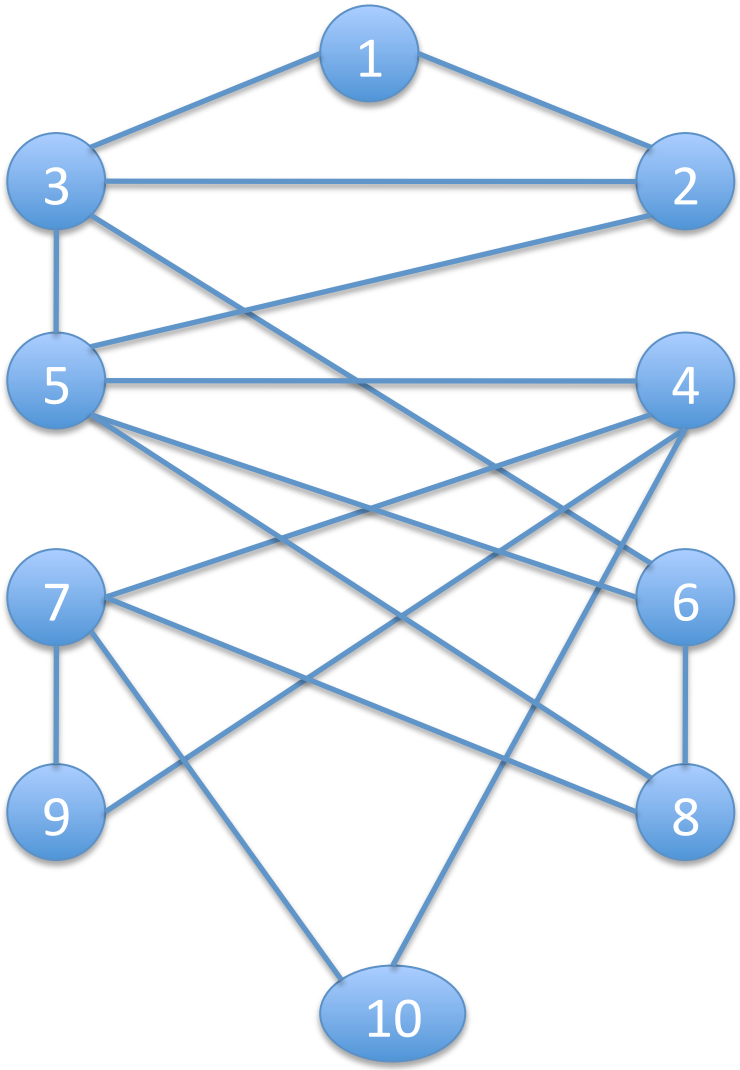
Use the greedy algorithm to find a minimum set of tagSNPs with a $r \geq 0.7$.

	1	2	3	4	5	6	7	8	9	10
1	1	0.9	0.85	0.5	0.4	0.2	0.2	0.15	0.15	0.1
2		1	0.95	0.5	0.8	0.2	0.2	0.15	0.15	0.1
3			1	0.65	0.9	0.7	0.5	0.5	0.3	0.2
4				1	0.85	0.5	0.85	0.6	0.7	0.7
5					1	0.75	0.6	0.75	0.6	0.5
6						1	0.6	0.75	0.4	0.3
7							1	0.8	0.85	0.8
8								1	0.6	0.5
9									1	0.5
10										1

HW2 Pr 2.2.1 – Finding Tag SNPs

Find the node with the most edges

Out degree count
1: 2
2: 3
3: 4
4: 4
5: 5 (highest)
6: 3
7: 4
8: 3
9: 2
10: 2



Tags 5

HW2 Pr 2.2.1 – Finding Tag SNPs

Find the node with the most edges

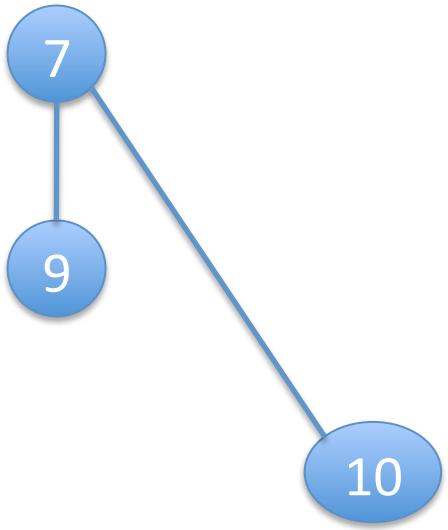
Out degree count

1: 0

7: 2

9: 1

10: 1



Tags 5, 7

HW2 Pr 2.2.1 – Finding Tag SNPs

Find the node with the most edges

Out degree count

1: 0

Tags 5, 7, 1

HW2 Pr 2.2.2 – Computing Power

2.2.2 Computing Power

Assume that the relative risk of one of tag SNPs in the greedy solution is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?

Again 4 steps !

Step 1. Compute p^+_A, p^-_A and p_A for each MAF using relative risk and MAF

-- We already computed this in problem 2.1

HW2 Pr 2.2.2 – Computing Power

Step 2. Compute NCP for each Tag SNP using p^{+}_{A}, p^{-}_{A}, p_A,N, and NCP for non-tagged SNP using NCP of Tag SNP and its correlation to Tag SNP

-- We already computed NCP for each Tag SNP in problem 2.1

-- NCP of non-tagged SNP is

-- Tag SNPs and correlated SNPs are

SNP 1: none

SNP 5: 2, 3, 4, 6, 8

SNP 7: 9, 10

-- NCP of correlated SNPs:

NCP of SNP 2 = NCP of SNP 5 * 0.8

NCP of SNP 3 = NCP of SNP 5 * 0.9

NCP of SNP 4 = NCP of SNP 5 * 0.85

NCP of SNP 6 = NCP of SNP 5 * 0.75

NCP of SNP 8 = NCP of SNP 5 * 0.75

NCP of SNP 9 = NCP of SNP 7 * 0.85

NCP of SNP 10 = NCP of SNP 7 * 0.8

-- Can re-use R code in Pr 2.1 like

c(0.8,0.9,0.85,0.75,0.75)* ncp(2.0,0.3,200)

HW2 Pr 2.2.2 – Computing Power

Step 3. Compute Power for each SNP using NCP and α (don't forget Bonferroni correction & the number of tag SNPs!)

-- Since there are 3 tags, $\alpha_S = \alpha/3 = 0.05/3 = 0.01666667$

$$\text{power at SNP } 1 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP1)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP1)})$$

$$\text{power at SNP } 2 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP2)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP2)})$$

$$\text{power at SNP } 3 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP3)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP3)})$$

$$\text{power at SNP } 4 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP4)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP4)})$$

$$\text{power at SNP } 5 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP5)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP5)})$$

$$\text{power at SNP } 6 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP6)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP6)})$$

$$\text{power at SNP } 7 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP7)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP7)})$$

$$\text{power at SNP } 8 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP8)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP8)})$$

$$\text{power at SNP } 9 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP9)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP9)})$$

$$\text{power at SNP } 10 = \Phi(\Phi^{-1}(0.05/3/2) + \text{NCP(SNP10)}) + 1 - \Phi(-\Phi^{-1}(0.05/3/2) + \text{NCP(SNP10)})$$

Need to modify R code for power in Pr 2.1

Step 4. Average Power to compute total power using power of each SNP

-- Average power of 10 SNPs

HW2 Pr 2.3 – Optimal algorithm

2.3.1 Finding Tag SNPs

The greedy solution for finding the minimum set of tag SNPs is not the optimal solution. What is the optimal solution?

2.3.2 Computing Power

Assume that the relative risk of one of tag SNPs in the optimal solution is 2.0 (we do not know which one). Assume that we are collecting 100 case and 100 control individuals. With $\alpha = 0.05$, what is the power of this association study?

Basically the same problem as Pr 2.2, but you need to find the optimal solution for Tag SNPs, and its power

HW2 Pr 3 – Indirect Association Study Problem

3.1 Calculating Correlation

Let's assume that we have a following reference dataset of 10 individuals representing a population such as the HapMap. What is the correlation, r, between SNP A and SNP B?

Individuals	SNP A	SNP B
Individual 1	A	A
Individual 2	a	a
Individual 3	A	A
Individual 4	A	a
Individual 5	a	a
Individual 6	A	A
Individual 7	a	A
Individual 8	A	A
Individual 9	A	a
Individual 10	A	A

HW2 Pr 3.1 – Calculating Correlation

The correlation equation is

$$\frac{p_{AB} - p_A p_B}{\sqrt{p_A(1-p_A)}\sqrt{p_B(1-p_B)}}$$

$p_A = 0.3$, $p_B = 0.4$, $p_{AB} = 0.2$

Or, you can use R to compute correlation. Encode A as 1 and a as 0 (reverse works too)

```
> snpA = c(1,0,1,1,0,1,0,1,1,1)
> snpB = c(1,0,1,0,0,1,1,1,0,1)
> cor(snpA,snpB)
```

HW2 Pr 3.2 – Indirect Association Power

Assume the causal SNP is B, but we collect SNP A. Assume that true case probability and and true control probability are 0.4 and 0.5 respectively at SNP B. If we collect 500 case and 500 control individuals and have a significance threshold of 0.05, what is the power at SNP A? (Note : Use the correlation that you get from above question)

First, calculate non-centrality parameter of SNP B

$P^{+}_{B} = 0.4$, $P^{-}_{B} = 0.5$, $P_{B} = (0.4+0.5)/2 = 0.45$, $N = 1000$

$$\lambda_B \sqrt{N} = \frac{p_B^+ - p_B^-}{\sqrt{2/1000}\sqrt{p_B(1-p_B)}} = \frac{0.4-0.5}{\sqrt{1/500}\sqrt{0.45*(1-0.45)}}$$

Second, calculate non-centrality parameter of SNP A

$$\lambda_A \sqrt{N} = r \cdot \lambda_B \sqrt{N}$$

Lastly, calculate the power using NCP of SNP A

$$\Phi(\Phi^{-1}(\alpha/2) + \lambda_A \sqrt{N}) + 1 - \Phi(-\Phi^{-1}(\alpha/2) + \lambda_A \sqrt{N})$$

HW2 Pr 4 – Association Study with Multiple Disease

(Grad Students ONLY)

We know from the homework, that the most efficient association studies have the same number of cases and controls. The Wellcome Trust Case Control Consortium used 2000 cases and 3000 controls for each of their disease associations. If you use the formula from the homework, this turns out to be equivalent to an balanced case/control study with 2400 each. So in essence, they used 5000 people but only got the equivalent power of using 4800.

However, what they did was have 7 diseases where they collected 2000 cases and they used the same 3000 controls for each association study. So they effectively used the 3000 controls many times while the each cases individual was only used once. They collected a total of 7*2000+3000=17000 individuals.

Now the question is did they collect the right number of cases and controls in this kind of scenario? If not, how many should they have collected. What if there were only 3 diseases (the total number of individuals is 3*2000+3000 = 9000)? How about 10 diseases (the total number of individuals is 10*2000+3000 = 23000)?

HW2 Pr 4 – Association Study with Multiple Disease

- The total number of individuals collected is 7*2,000 + 3,000 = 17,000 individuals

- The question is, did they collect the right number of cases and controls in this scenario under the assumption that the number of cases is the same for all 7 diseases and the total number of individuals they collect is 17,000?

- In other words, does collecting 2,000 cases for each disease and collecting 3,000 controls maximize the power given the constraint that we collect 17,000 individuals?

- For example, what if we collect 1,500 cases for each disease (7 * 1500 = 10,500) and collect 6,500 controls (17,000 – 10,500 = 6,500). Does this have higher power than collecting 2,000 cases and 3,000 controls?

- If not, how many should they have collected?

- What if there were only 3 diseases, 10 diseases?

HW2 Pr 4 – Association Study with Multiple Disease

In the unbalanced study, remember that NCP is $\lambda_A \sqrt{\frac{2(N^+N^-)}{N^+ + N^-}}$

λ_A does not depend on N^+ or N^-, so we want to know the value of N^+ and N^- that maximzes the power

There are several ways for finding the value, and one way is taking derivative

$N^+ + N^- = 2N$, where N^+ is # of case chromosomes, N^- is # of control chromosomes, N is the total # of individuals, and we have 17,000 total individuals, so 34,000 total chromosomes

$$34000 = 7N^+ + N^-$$

$$N^- = 34000 - 7N^+$$

$$\sqrt{\frac{2N^+N^-}{N^+ + N^-}} = \sqrt{\frac{2N^+(34000 - 7N^+)}{N^+ + 34000 - 7N^+}} = \sqrt{\frac{68000N^+ - 14N^{+2}}{34000 - 6N^+}}$$

$$\left(\sqrt{\frac{68000N^+ - 14N^{+2}}{34000 - 6N^+}}\right)\frac{d}{dN^+} = \frac{\text{numerator}}{\text{denominator}}$$

Set the numerator equal to 0, then solve for N^+, then you can solve for N^- using $N^- = 34000 - 7N^+$

Hint : You can use online math tool (e.g. WolframAlpha) to compute the derivative and to solve N^+

HW2 Pr 4 – Association Study with Multiple Disease

If there are 3 diseases, then we have $3*2000 + 3000 = 9000$ total individuals. So,

$$18000 = 3N^{+} + N^{-}$$

$$N^{-} = 18000 - 3N^{+}$$

If there are 10 diseases, then we have $10*2000 + 3000 = 23000$ total individuals. So,

$$46000 = 10N^{+} + N^{-}$$

$$N^{-} = 46000 - 10N^{+}$$